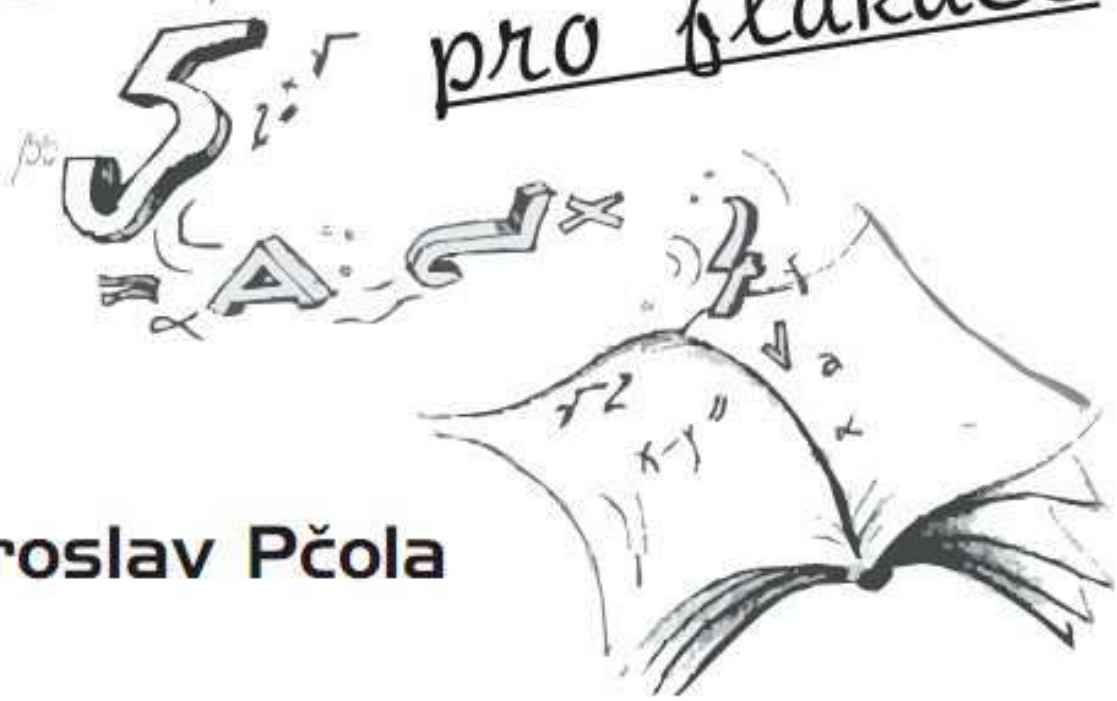
Statistika

pro fláčače



Jaroslav Pčola

Obsah

Opravy pro toto vydání: opravy2.proflakace.cz

Úvod

Mojou pracovnou metódou je neučiť sa, a potom, keď sa všetko nakopí, v strese to zvládnuť. Nevravím, že je to metóda správna, ale u mňa funguje. Keď som sa v roku 2007 začal učiť na skúšku zo štatistiky, už v prvý deň som zistil, že mám problém. Nerozumel som tomu, nevedel som z ktorého konca začať, bolo toho veľa a ja som nemal absolútne žiaden základ. Po čase som sa však do toho dostal a pochopil, že to vôbec nie je až také zložité, ako sa spočiatku zdalo. Pochopil som tiež, že ak by som na začiatku mal literatúru šitú pre mňa, pre flákača, ušetril by som si niekoľko veľmi trpkých dní. Tak vznikol môj záväzok, v prípade, že 4ST201 zvládnem, takéto skriptum napísať.

Z týchto pohnútok vznikla prvá Štatistika pre flákačov, čo bolo skriptum robené na kolene, hotové za pár dní. Veľmi som sa nezamýšľal nad tým, či tam mám alebo nemám chyby, či to, ako vnímam teóriu ja, je skutočne pravda. Podstatné bolo, že to fungovalo, dalo sa z toho učiť, typové príklady tam boli a naštartovalo to človeka do ďalšej aktivity. Zosmolil som to asi za týždeň a samozrejme to bolo na nete (a je a bude) zadarmo. Český preklad, ktorý sa potom objavil, prekrútil síce dosť veľa mojich viet, čím vniesol do matroša ešte viac chybovosti, ale vďaka nemu som pochopil, že to jednoducho musí byť česky.

Dlho som sa potom odhodlával k dopísaniu materiálu. Vedel som, že ak sa do toho pustím, už sa neuspokojím s takým amatérskym spracovaním a tým pádom mi bolo jasné, že to bude veľa práce. Dnes, keď píšem túto stránku, ktorú som si nechal ako poslednú, už viem, že som to podcenil. Je to strašne veľa práce ☺. Musel som sa do štatistiky znova dostať, naštudovať si nové požiadavky, počítať príklady a nakoniec aj formátovať, vybavovať, prepisovať, kontrolovať a tak ďalej. Napriek tomu netvrdím, že toto, čo práve držíte v rukách je bez chyby. Mohol som niečo pochopiť zle, mohol som sa pomýliť vo výpočte a iste nájdete aj gramatické chyby. Stále to nie je substitút k učebnici, nie je to stostránkový materiál, ako som si ho na začiatku predstavoval. Ale myslím, že taký materiál ani nikto nepotrebuje. To čo držíte v rukách vás nakopne a viem si predstaviť, že ak ho prejdete celý za jeden deň a predtým ste o štatistike nepočuli ani v autobuse, je reálna šanca, že na druhý deň ten test napíšete so slušným výsledkom. A práve to má byť cieľom Štatistiky pre flákačov, poskytnúť zhustené informácie, očistené od balastu a podané tak, aby tomu pochopil každý. Aj za cenu nepresnosti, neúplnosti, či zjednodušenia. Ospravedlňujem sa preto vopred za chyby ktoré sa tu môžu nachádzať, bol som na to sám a nemal som žiadneho konzultanta, ak by ste však mojim konzultantom chceli byť, budem veľmi rád ak ma na chyby upozorníte mailom a ja sa ich budem snažiť uviesť na pravú mieru na svojej internetovej stránke.

Dostal som sa teda až k tlačenej verzii. Tlačená preto, lebo papier je proste papier a kniha vonia, a preto, že stále neexistuje použiteľný systém ako predávať texty na internete v komunite, kde sa všetko zdieľa. A predaj bol nutnou voľbou, ak som chcel dosiahnuť takúto kvalitu, zaplatiť preklad, učebnice, investovať viac než mesiac každodennej práce. Nečakám, že na tomto zarobím, cena je príliš nízka a náklady vysoké, bojím sa kopírovania či skenovania, ktoré by ma asi knokautovalo a moju prácu ohodnotilo ako bezcennú. Robím to hlavne pre dobrý pocit, preto, že som vždy chcel napísať knihu, preto, že sa neskutočne teším z každej referencie na borcovi a napĺňa ma radosťou, že vďaka mojej práci sa vyhnú útrapám učenia sa štatistiky stovky ľudí.

Prajem vám teda, aby ste sa pri čítaní tohto materiálu bavili, aby ste mali radosť, ako rýchlo štatistiku zvládate a aby výsledok vašej skúšky bol našim úspechom. Našim spoločným cieľom môže byť zvýšenie úspešnosti predmetu 4ST201 nad 80%. Verím, že sa nám to podarí. Veľa síl a šťastia.

Jaro

„Elementary, my dear Watson."

1 Základní pojmy

Největším nepřítelem procesu učení se (kromě vlastní neochoty) je jednoznačně nepochopené slovo. Pokud se opravdu chcete statistiku naučit a text skutečně vnímat, nesmíte přejít bez povšimnutí slovo, kterému nerozumíte, ale naopak – musíte zjistit jeho význam. Pokud ho jen tak přeletíte, je sice možné, že text pochopíte, ale s mnohem větší pravděpodobností se probudíte za 10 řádků s tím, že nevíte, o čem čtete. Já se vám to budu snažit ještě ulehčit, a to tak, že každý pojem, který tu budu používat vám podrobně a srozumitelně definuji – moje oblíbené slovo je *například*.
Tento úvod by měl definovat základní pojmy, se kterými ve statistice pracujeme. Samotná statistika je věda, která se zjednodušeně řečeno zabývá jednak pravděpodobností a na druhé straně tím, že analyzuje velké množství shromážděných dat (údaje o počasí za poslední dva roky, průzkum o nanucích na vzorku 2000 osob). Tato data třídí, hledá mezi nimi souvislosti, vyjadřuje se k pravděpodobnému budoucímu vývoji a tak podobně. Díky statistice dokážeme na základě velkého množství nic neříkajících čísel vyslovit nějaký závěr, který může mít pro nás větší či menší užitek. Například předpovědi počasí na období vzdálenější než čtrnáct dní jsou prakticky založené na statistických údajích o počasí a historických křivkách teplot a jen minimálně na aktuální situaci v atmosféře.
Ve statistice tedy pracujeme s opravdu velkým množstvím dat (statistický úřad se například zabývá většinou zkoumáním celého národa nebo celého průmyslového odvětví), většinou ale zkoumáme stejné znaky, které má mnoho stejných subjektů – například příjem domácností, zisk podniků chemické výroby atp. Subjekt, který zkoumáme, nazýváme **statistickou jednotkou** (to je ta domácnost, podnik) a jeho znak, vlastnost nazýváme **statistický znak** (příjem, zisk, velikost atp.). Tyto znaky mohou mít různý charakter, pro začátek mohou být **číselné nebo slovní**. Ty číselné se potom liší podle toho, zda je možné, aby nabývaly jakékoli hodnoty v určitém rozsahu (například chleba může vážit od 0 gramů přes 124,5432 gramů až do jakéhokoli čísla) – ty jsou **spojité**, nebo nabývají jen určité hodnoty (počet dětí jen 1, 2, 3, 4...; výsledek zápasu len 0, ½, 1) – ty jsou **nespojité**. Slovní znaky zase dělíme na **nominální,** u kterých neumíme určit, co je lepší, resp. co je víc (jako například místo bydliště - Praha, Brno, Olomouc – nedá se určit pořadí) a na **pořadové**, kde můžeme určit nějaké pořadí (známka ve škole ve slovním vyjádření – dobrý je určitě lepší, než nevyhověl).

1.1 Třídění dat

Pokud bychom si vzali nějaká konkrétní data, například výšky žáků ve třídě, které, vzhledem k tomu, že se na ně díváme jako statistici, budeme nazývat **statistický soubor**, vypadaly by asi nějak takto:

150, 160, 156, 175, 148, 175, 181, 150, 175, 189, 179, 156, 181, 179, 176

Z takové nezpracované informace toho moc nezjistíme, proto je vhodné si tyto údaje utřídit. V mnoha případech je vhodnou formou třídění **tabulka rozdělení četnosti**, kde si pod sebe napíšeme, jaké údaje jsme naměřili (seřadíme je podle velikosti) a kolikrát:

148	1
150	2
156	2
160	1
175	3
...	...

Čísla udávající počet opakování nějaké hodnoty nazýváme četnost. Kdybychom k tomuto nominálnímu vyjádření dopočítali i procentuální (výška 175 byla v souboru 3krát – to je nominální vyjádření a žáci výšky 175 byli 3 z 15ti = 3/15 = 0,2 = 20% je procentuální vyjádření), nazývali bychom jej **relativní četností**.

V tomto případě však vidíme, že tabulka rozdělení četnosti asi není nejvhodnější, pokud máme mnoho údajů, které se od sebe liší. Použijeme proto **intervalové rozdělení četnosti**, což znamená, že si rozdělíme naměřené hodnoty do intervalů a až pak vyjádříme, kolikrát jsme jakou hodnotu naměřili. Velikost intervalů se může určit pomocí některých matematických metod, ale není žádný zločin, když si intervaly určíme jen tak od oka, podle toho, co nám našeptává naše logika. Samozřejmě by bylo vhodné, aby měly intervaly stejnou velikost.

<140 – 150)	1
<150 – 160)	4
<160 – 170)	1
<170 – 180)	6
<180 – 190)	3

Další možností pro třídění dat ne příliš vhodnou pro další počítání s naměřenými čísly, ale vysoce efektivní při prezentování výsledků měření, jsou grafy.

1.2 Míry úrovně – polohy

Když si představíme číselnou osu a chtěli bychom na ní naznačit to velké množství našich čísel *(představme si např. výšky dětí)* jen jediným bodem, poloha je to místo, kam bychom dali tečku. Tedy – je to nějaká střední hodnota. Možností, jak ji určit, máme vícero.

Průměr – průměrů existuje několik, nás zajímá hlavně aritmetický, takže klasicky sčítám všechny hodnoty, co jsem naměřil, a vydělím to počtem měření. Značit jej budeme $\overline{x}$. V některých konkrétních případech ale musíme využít i průměr harmonický – například při počítání příkladů o průměrné rychlosti auta, v jiných případech i průměr geometrický.

Medián – je hodnota středního členu. Kdybychom postavili děti podle výšky vedle sebe, výška toho, kdo bude uprostřed, bude medián. Pokud by byl sudý počet, medián by byl průměrem dvou hodnot ve středu. Označuje se $\tilde{x}$.

Modus – je nejčastěji se vyskytující hodnota, takže z čísel 2,2,3,4,5,5,5,6 by byl modus 5, protože je tam třikrát. Značíme ho $\hat{x}$.

1.3 Míry variability

Variabilita je vzdálenost našich naměřených hodnot od střední hodnoty. Můžeme ji vyjádřit různě, např. máme patnáct žáků a průměr jejich výšek je 168,66. Tím jsme vyjádřili polohu, a když řekneme, že jejich výšky se pohybují od 148 do 189, tak jsme vyjádřili variabilitu. Konkrétně takové vyjádření se nazývá **variační rozpětí**, kdy jednoduše odečteme nejmenší hodnotu od největší. Tzn. 189-148 = 41. Tento způsob je ale velmi náchylný na extrémy. Kdybychom měli jen jediného 130 cm vysokého trpaslíka, naše variační rozpětí by bylo 59. Proto moudří lidé vymysleli **rozptyl**. Počítáme-li tento rozptyl, tak od každé naměřené hodnoty odečteme průměr všech naměřených hodnot, vyjde nám tedy odchylka od střední hodnoty. Tu potom umocníme na druhou, abychom neměli záporná

čísla. Když všechny tyto odchylky umocněné na druhou sečteme a poté vydělíme jejich počtem (zprůměrujeme), výsledkem je rozptyl. Označujeme jej:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$$

Rozptyl nám však toho o skutečné variabilitě moc nepoví, protože všechny odchylky jsou v něm umocněné na druhou, takže při:

Průměr = 168,66
Odchylky na druhou: $(150\text{-}168{,}66)^2 + (160\text{-}168{,}66)^2 + (156\text{-}168{,}66)^2 + \ldots = 3092{,}24$
Výsledek vydělíme patnácti (počet hodnot = n) a máme rozptyl 206,15 cm na druhou.

Hodnota v centimetrech na druhou (nebo v jakékoli jiné jednotce na druhou) ale není příliš srozumitelná, proto se používá odmocnina z rozptylu, která má název **směrodatná odchylka**, v našem případě odmocnina z 206,15 je 14,36 centimetrů. Toto číslo přibližně vyjadřuje, že „více než 50% naměřených hodnot (výšek) se neodchyluje od průměru v obou směrech o více než 14,35 centimetru". Směrodatnou odchylku značíme stejně jako rozptyl, jenom ne na druhou - s_x. Vzorců na výpočet rozptylu existuje několik, ale jednoznačně nejpoužívanější pro běžné počítání je tzv. **výpočetní tvar rozptylu**. Jeho vzorec je matematickým zjednodušením původního vzorce a postup výpočtu rozptylu spočívá v tom, že zprůměrujeme druhé mocniny naměřených hodnot a od toho odečteme průměr hodnot umocněný na druhou:

$$s_x^2 = \overline{x^2} - \overline{x}^2$$

Variabilitu je možné definovat i relativně, takže ne v nějakých jednotkách (jako v případě směrodatné odchylky), ale jen bezrozměrným číslem nebo procentem. K tomu slouží **variační koeficient**, který je poměrem směrodatné odchylky a aritmetického průměru.

$$V_x = \frac{s_x}{\overline{x}}$$

Hodnota variačního koeficientu se může pohybovat od $-\infty$ do $+\infty$ a říká nám, nakolik je náš soubor hodnot nesourodý (nakolik jsou ty čísla, které tvoří soubor, rozházené po číselné ose). Hodnoty od -0,5 do +0,5 ukazují, že čísla se pohybují v blízkosti střední hodnoty, a naopak ostatní hodnoty svědčí o větší nebo menší nesourodosti (čím větší číslo do plusu nebo mínusu, tím je soubor méně sourodý).

1. *Z tabulky rozdělení četnosti vypočítejte průměr, rozptyl a směrodatnou odchylku věku 15ti pracovníků firmy:*

Věk	33	34	37	41	43	47	58
Počet	3	2	2	4	2	1	1

a. Průměr

Sečteme všechny hodnoty a vydělíme 15ti. Pozor na to, že se v souboru mnohé hodnoty nacházejí vícekrát, proto to musíme zohlednit i v čitateli:

$$\overline{x} = \frac{3 \cdot 33 + 2 \cdot 34 + 2 \cdot 37 + 4 \cdot 41 + 2 \cdot 43 + 1 \cdot 47 + 1 \cdot 58}{15} = \frac{596}{15} = 39,73$$

b. rozptyl

Uděláme součet druhých mocnin, potom je vydělíme 15ti a odečteme druhou mocninu již vypočítaného průměru:

$$\overline{x^2} = \frac{3 \cdot 33^2 + 2 \cdot 34^2 + 2 \cdot 37^2 + 4 \cdot 41^2 + 2 \cdot 43^2 + 1 \cdot 47^2 + 1 \cdot 58^2}{15} = \frac{24312}{15} = 1620,8$$

$$\overline{x}^2 = 39,78^2 = 1578,47$$

$$s_x^2 = 1620,8 - 1578,47 = 42,33$$

Rozptyl je 42,33 roků na druhou.

c. směrodatná odchylka

Je odmocninou z rozptylu, tedy odmocnina z 42,33. Směrodatná odchylka je 6,5 roku.

2. *Bezdomovec na Václaváku vyžebral za jeden den od 115ti lidí celkem 1400 korun a rozptyl příspěvků byl 140. Jeho kolega na Hlavním nádraží získal od 86ti lidí 1300 korun a rozptyl jeho příspěvků byl 90. Který bezdomovec dosáhnul ten den větší relativní variability příspěvků?*

Směrodatnou odchylku do vzorce získáme odmocněním rozptylu a průměr vydělením sumy vyžebraných peněz počtem lidí, kteří přispěli.

$$V_1 = \frac{\sqrt{140}}{1400/115} = 0,972 \qquad V_2 = \frac{\sqrt{90}}{1300/86} = 0,628$$

Větší variability příspěvků dosáhnul bezdomovec na Václaváku.

3. *Z 10 hodnot byl vypočítán průměr 100 a rozptyl 200. Dodatečně jsme však zjistili, že každá hodnota byla o 10 nadhodnocená. Jaké jsou skutečné hodnoty průměru a rozptylu?*

Pokud 10 hodnot má průměr 100, tak součet těch hodnot je 10x100=1000. Každá hodnota je v skutečnosti o 10 nižší, což nám při deseti hodnotách udělá 10x10 = 100. Celkový součet měl být o 100 nižší = 1000-100 = 900. Takže nový průměr je 900/10 = 90.

Abychom zjistili novou hodnotu rozptylu, vůbec nemusíme počítat. Pokud si uvědomíme, že rozptyl říká, nakolik jsou naše čísla roztroušené po číselné ose – neboli jak velké jsou vzdálenosti mezi nimi – je úplně jedno, jestli jsou ty čísla 200, 300 a 3000, anebo 100, 200 a 2900. Proto když **všechna čísla** zvětšíme nebo zmenšíme o libovolnou konstantu, rozptyl zůstává stále stejný. V našem případě zůstává i nový rozptyl 200.

4. *Auto projelo celou trať o délce 100 km průměrnou rychlostí 84,25 km/h. V prvním úseku o délce 30 km dosáhlo průměrné rychlosti 70 km/h, ve druhém úseku o délce 15 km dosáhlo průměrné rychlosti 50 km/h. Jakou mělo auto průměrnou rychlost na třetím úseku trati?*

Pokud počítáme průměr průměrné rychlosti, **musíme použít harmonický průměr**, to je asi jediná záludnost tohoto příkladu. V případě, že jednotlivé části trati jsou nestejně dlouhé, používáme vzorec pro vážený harmonický průměr. Každý jednotlivý kilometr trasy vystupuje jako jedno n – suma n je tedy 100. Ve jmenovateli potom dosazujeme průměrnou rychlost – x a počet kilometrů, na kterých této rychlosti auto dosahovalo – n.

$$\bar{x}_H = \frac{\sum n_i}{\sum \frac{n_i}{x_i}} \qquad \sum n_i = 100 \qquad \bar{x}_H = 84,25 \qquad x_3 = ?$$

$$84,25 = \frac{100}{\frac{30}{70} + \frac{15}{50} + \frac{55}{x_3}}$$

$$84,25 \cdot \left(\frac{30}{70} + \frac{15}{50} + \frac{55}{x_3} \right) = 100$$

$$84,25 \cdot \frac{55}{x_3} = 100 - 84,25 \cdot \frac{30}{70} - 84,25 \cdot \frac{15}{50} = 38,61785$$

$$x_3 = \frac{55}{38,61785 / 84,25} = 119,99\, km/h$$

Pokud bychom to počítali aritmetickým průměrem, vyjde nám 100 km/h. To ale není správný výsledek.

2 Počet pravděpodobnosti

Všechny děje, které se ve světě odehrávají, mají svoji pravděpodobnost. Základní **vlastností pravděpodobnosti je, že se pohybuje v rozmezí 0 až 100 procent**. Když zkoumáme pravděpodobnost nějakého jevu za daných podmínek, například pravděpodobnost srážky Země s Jupiterem za 15 sekund, tak jeho nulová pravděpodobnost značí, že daný jev nemůže nastat, je to **jev nemožný**. Naopak stoprocentní pravděpodobnost říká, že jev určitě nastane – je to **jev jistý**. Většina věcí, které zkoumáme, ale nemá charakter jistých nebo nemožných jevů, například na hrací kostce nám může padnout 6 čísel, a že nám padne právě trojka, má určitou (ne nulovou ani stoprocentní) pravděpodobnost. Padnutí trojky na hrací kostce proto nazýváme **jevem náhodným**. To, že kostkou házeme a sledujeme, jestli nám trojka padne nebo ne, nazýváme zase **náhodným pokusem**.
Jako teoretický základ počítání s pravděpodobnostmi používáme dvě definice pravděpodobnosti. První se nazývá **klasická definice** a zakládá se na tom, že pravděpodobnost každé varianty náhodného pokusu je stejná. To platí například u hrací kostky, kde každé číslo by mělo padat stejně často, anebo při tahání čísla z osudí, kde se každé nachází pouze jednou, a tedy je stejně pravděpodobné vytáhnutí jakéhokoli čísla a tak podobně. V tom případě umíme vypočítat pravděpodobnost jako **poměr počtu nám příznivých (námi očekávaných) variant jevu a počtu celkových možných variant**. Tedy pokud potřebujeme trojku na kostce, pravděpodobnost, že nám padne, je:

1 (počet příznivých variant) děleno 6 (počet možných) = 1/6

Naopak **statistická definice** pravděpodobnosti se používá, když není splněný předpoklad stejných pravděpodobností jednotlivých variant, například při autohavárii, ve které se zraní přesně jeden člověk, není pravděpodobnost zranění stejná pro všechny místa v autě. Některé jsou rizikové více a některé méně. V takovém případě je možné určit **pravděpodobnost na základě údajů za dlouhé období** o daném jevu, respektive na základě experimentálních měření a po jejich zhodnocení dospět k nějaké pravděpodobnosti, se kterou daný jev nastává (pravděpodobnost, že se zraní řidič, pravděpodobnost, že se zraní spolujezdec atd.). Tato pravděpodobnost je tím přesnější, čím víc údajů máme, ale nikdy není úplně přesná, je to vždy jen přibližná hodnota, okolo které se skutečnost pohybuje. V příkladech pravděpodobnost na základě statistické definice nepočítáme, ostatně by to ani nebylo možné, vždy ji dostaneme zadanou (např. poruchovost auta je 5%) a její hodnotu používáme pro další počítání s pravděpodobnostmi.

2.1 Průnik a sjednocení jevů

Máme-li více náhodných pokusů (například dva hody kostkou) a chceme určit pravděpodobnost, že nastane nějaká kombinace náhodných jevů (například na první kostce padne trojka a na druhé šestka), musíme použít vzorce pro počítání s pravděpodobnostmi. V zásadě jsou důležité jen dva případy a to:

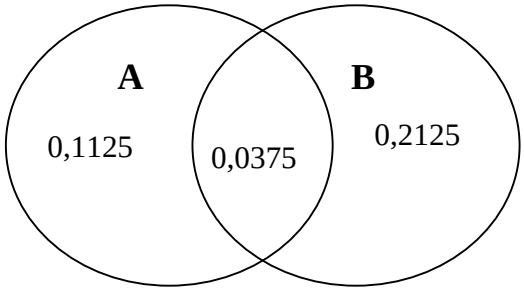
Průnik jevů = $P(A \cap B)$

Chci zjistit pravděpodobnost, že zítra ráno po dnešní chodbovici půjdu do školy na hodinu v 7:30. Řekl jsem si, že když mě vzbudí budík a probudím se ve své posteli, tak půjdu. Pravděpodobnost, že mě vzbudí budík je 0,7 a pravděpodobnost, že se probudím ve své posteli je 0,55. Pravděpodobnost, že do školy půjdu, je pravděpodobností průniku dvou jevů a vyjadřuje nakolik je pravděpodobné, že oba jevy nastanou **současně**. Ptám se na průnik, protože **jedině v případě, že nastanou oba jevy současně**, půjdu do školy. Hodnotu této pravděpodobnosti vypočítáme jako součin pravděpodobností jevů, které mají nastat současně**:** $P(A \cap B) = P(A) \cdot P(B)$. V našem případě tedy P (půjdu do školy) = 0,7 x 0,55 = 0,385 = 38,5.

Sjednocení jevů = $A \cup B$
Pravděpodobnost, že neudělám zkoušku z matematiky, je 0,15. Pravděpodobnost, že neudělám zkoušku ze statistiky, je 0,25. Stačí, když neudělám jeden z těchto předmětů, a vyhodí mě ze školy. Pravděpodobnost, že mě vyhodí ze školy, je pravděpodobností sjednocení = $P(A \cup B)$. Slovně to můžeme vyjádřit: „Pravděpodobnost, že neudělám zkoušku z matematiky **nebo** zkoušku ze statistiky". Sjednocujeme tyto dva jevy, protože **je jedno, který z nich nastane, nebo zda nastanou současně**, každopádně to pro mě bude znamenat vyloučení ze školy. Pro vypočítání pravděpodobnosti sjednocení těchto jevů použijeme vzorec:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

tedy jednoduše sečteme pravděpodobnosti jevů a odečteme pravděpodobnost jejich průniku, neboli situaci, kdy nastanou oba jevy současně. Proč ji musíme odečíst, nejlépe ilustruje diagram:



V bublině A je pravděpodobnost, že neudělám matematiku, složená ze dvou pravděpodobností – první, že neudělám pouze matematiku, a druhé, že neudělám matematiku a zároveň i statistiku = 0,1125 + 0,0375 = 0,15. Obdobně v bublině B máme 0,2125+0,0375 = 0,25. Jejich průnik (pravděpodobnost, že neudělám ani jednu zkoušku) jsem připočetl k jednomu i ke druhému jevu, ovšem **díky tomu, že se na oba jevy díváme jako na celek, nemůžeme přičítat neudělání statistiky i matematiky k oběma bublinám – tedy dvakrát**, ale musíme jej počítat pouze jednou. Musíme proto tento průnik od výsledku P (A) + P (B) jednou odečíst.

Je to obdobné, jako když sbírám kartičky hokejistů, a mám sto různých, kamarád má také sto různých, takže je dáme dohromady a budeme mít společnou sbírku. Jestliže však později zjistím, že mám 20 stejných kartiček jako kamarád, tak dohromady máme jen A+B - $A \cup B$ *= 100+100-20=180 různých kartiček.*

Pravděpodobnost mého vyhození ze školy je tedy 0,15 + 0,25 - 0,0375 = 0,3625. Průnik samozřejmě **nemusím odečítat, pokud žádný neexistuje**. Např. nemůže být -10 stupňů a zároveň pršet.

1. *V pytlíku máme 10 černých a 5 bílých kuliček, jaká je pravděpodobnost, že vytáhneme bílou?*

Každá kulička v pytlíku má stejnou šanci na vytáhnutí, výběr každé kuličky je stejně pravděpodobný. Ptáme se na pravděpodobnost vytáhnutí bílé – její vytáhnutí bude pro nás příznivý jev. Počet příznivých výsledků je proto 5 (a naopak máme 10 nepříznivých výsledků – vytáhnutí jakékoli z deseti černých kuliček) a počet všech kuliček je 15. Proto výsledná pravděpodobnost je 5/15 = 1/3.

2. Jaká je pravděpodobnost, že na kostce padne při prvním hodu dvojka a při druhém šestka?

Jsou to dva nezávislé jevy, každé číslo při každém hodu padá s pravděpodobností 1/6. Pravděpodobnost *padnutí dvojky v prvním hodu* (jev A) je proto 1/6 a *padnutí šestky ve druhém hodu* (jev B) je též 1/6. My chceme, aby nastal první i druhý jev současně. Slovo současně indikuje, že počítáme průnik:

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

Pravděpodobnost tohoto jevu je 1/36.

3. Dvě prasátka zůstaly samy doma. Vlk přišel za nimi zahrát si poker a bouchá na dveře. Prasátko Bambi s pravděpodobností 0,2 poslouchá metal a vlka neuslyší, prasátko Cecil s pravděpodobností 0,4 hraje ve sklepě na bicí a v tom případě vlka taky neuslyší. Prasátka navíc s pravděpodobností 0,5 hrají spolu Counter Strike a vůbec neregistrují okolí. Jaká je pravděpodobnost, že si ani jedno prasátko nevšimne, že vlk bouchá na dveře?

Jestliže hrají hru (jev A), což je pravděpodobné na 50%, ani jedno si ho nevšimne. Jestliže nehrají, tak si ho ani jedno nevšimne s pravděpodobností rovnou průniku dvou jevů – prasátko Bambi jej neuslyší (jev B) a současně i prasátko Cecil (jev C). Pravděpodobnost tohoto průniku je rovna součinu 0,2 x 0,4 = 0,08. To, že si prasátka vlka nevšimnou, znamená, že buď hrají hru, nebo nastala situace, že se věnují metalu a bubnům. Spojka nebo nám naznačuje, že tuto pravděpodobnost vyjádříme pomocí sjednocení 0,5 + 0,08 = 0,58. Neodčítáme žádný průnik, protože není možné, aby prasátka hrály hru a zároveň se věnovaly bubnům a metalu. Prasátka nezaregistrují vlka s pravděpodobností 58%.

$$P = P(A) + P(B \cap C) = 0,5 + (0,2 \cdot 0,4) = 0,58$$

4. Házeme kostkou, dokud nám nepadne šestka. Nakolik je pravděpodobné, že se tak stane právě v šestém hodu?

Pravděpodobnost padnutí šestky je v každém hodu stejná – 1/6. Pravděpodobnost, že šestka nepadne, je 1 mínus 1/6 – teda 5/6. Dá se to spočítat taky opačně – 5/6 je pravděpodobnost, že nám padne jakékoli jiné číslo, takže když jiných čísel než je šestka je dohromady 5 a padnutí každého je pravděpodobné na 1/6, jejich suma bude 5/6. Pravděpodobnost, že 5x šestka nepadne a po šesté padne je průnikem šesti jevů. Řešením proto bude výsledek součinu: $\frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = 0,067$

5. Jaká je pravděpodobnost, že dvěma hody padne:
   a. jedna šestka
   b. ani jedna šestka
   c. dvě šestky

Pokud má padnout jedna šestka ve dvou hodech, tak může padnout buď v prvním, nebo ve druhém. Jedná se o sjednocení dvou jevů, přičemž oba jevy jsou zase průnikem dvou jevů.

P (jev A) = P (v prvním hodu padne 6, ve druhém ne) = 1/6 x 5/6 = 5/36
P (jev B) = P (v prvním hodu nepadne 6, ve druhém ano) = 5/6 x 1/6 = 5/36

$$P(A \cup B) = P(A) + P(B) = \frac{5}{36} + \frac{5}{36} = \frac{10}{36}$$

Jestliže ani jedna šestka nemá padnout, pravděpodobnost je 5/6 x 5/6 = 25/36
A nakonec pravděpodobnost padnutí dvou šestek je 1/6 x 1/6 = 1/36
Pozoruhodné je, že součet výsledků, které jsme vypočítali (10/36+25/36+1/36), se rovná jedné. Je to tak proto, že ze dvou hodů kostkou nemůžeme dostat nic jiného než 0, 1, nebo 2 šestky, a to s pravděpodobností, jakou udávají naše výsledky.

6. Z dvaceti lístků je 10 bílých. Taháme z osudí 3 lístky bez vracení. Jaká je pravděpodobnost, že vytáhneme tři bílé?

V takovém případě je v zásadě jedno, jestli taháme ty tři lístky zároveň anebo postupně. Můžeme tedy říct, že jde o pravděpodobnost, že nastane průnik tří náhodných jevů, přičemž první je vytáhnutí prvního bílého lístku (jeho pravděpodobnost je 10/20), potom druhého (9/19, protože už máme o jeden bílý lístek v osudí méně) a nakonec třetího (P = 8/18). Dáme to do součinu a dopočítáme:

$$P(A \cap B \cap C) = \frac{10}{20} \cdot \frac{9}{19} \cdot \frac{8}{18} = \frac{720}{6840} = 0{,}1053 .$$

O hodně hezčí a ve složitějších příkladech využitelnější metodou je ale použití kombinačního čísla. **Kombinační číslo** nám udává, kolika způsoby je možné vybrat prvky z nějaké větší množiny n prvků. Při používání kombinačního čísla upouštíme od představy, že taháme papírky postupně, ale představme si, že je taháme zároveň. Potřebujeme 3 bílé a dohromady je jich v osudí 10. Nyní si představme, že ty bílé papírky jsou označené od 1 do 10. Kolik možných trojic můžeme vytáhnout? Šlo by teoreticky začít je vypisovat na papír (123,124,125,126…), ale bylo by to velmi pracné. Efektně se to dá vypočítat pomocí kombinačního čísla $\binom{n}{k} = \frac{n!}{(n-k)!k!}$, kde *n* je velikost skupiny, ze které vybíráme = 10 prvků a *k* je velikost výběru = 3 prvky. Výpočet potom vypadá takto:

$$\binom{10}{3} = \frac{10!}{(10-3)!3!} = \frac{10.9.8.7!}{7!3!} = \frac{10.9.8}{3.2.1} = \frac{720}{6} = 120$$

Takže je 120 možných trojic bílých papírků a to je těch 120 trojic, jejichž vytáhnutí z celku 20ti papírků je pro nás příznivé. Ještě musíme zjistit počet všech možných trojic, které se dají z 20ti papírků vytáhnout, a když tato dvě čísla dáme do poměru podle vzorce klasické pravděpodobnosti, zjistíme pravděpodobnost vytáhnutí trojice bílých papírků.

$$\binom{20}{3} = \frac{20!}{(20-3)!3!} = \frac{20.19.18.17!}{17!3!} = \frac{20.19.18}{3.2.1} = \frac{6840}{6} = 1140$$

Jak vidíme, dospěli jsme ke stejným číslům jako při použití předchozí metody, pouze jsou vydělené šesti. To ale nic nemění na tom, že jejich poměr je 120/1140 = 0,1053.

7. Máme 15 červených a 25 modrých kuliček v pytlíku. Jaká je pravděpodobnost, že při tahání dvou kuliček bez vracení bude jedna z vytáhnutých červená a druhá bude modrá?

Ptáme se na dvojici, kde bude jedna červená a jedna modrá kulička. Počet možností, kterými se dá vytáhnout 1 červená z 15ti je právě 15 (můžete si to ověřit přepočítáním kombinačního čísla) a jedna modrá se dá z 25ti vytáhnout taktéž 25ti způsoby. Jestliže ke každé červené kuličce můžeme vytáhnout libovolnou modrou kuličku, počet všech dvojic, kde je jedna červená a jedna modrá kulička, bude 15x25 = 375. Kolik libovolných dvojic můžeme vytáhnout ze všech 40ti kuliček nám řekne kombinační číslo:

$$\binom{40}{2} = \frac{40!}{(40-2)!2!} = \frac{40.39.38!}{38!2!} = \frac{40.39}{2} = 780$$

Počet příznivých dvouprvkových výběrů je tedy 375 a celkový počet možných je 780. Pravděpodobnost vytáhnutí dvojice červené a modré kuličky je 0,48.

8. Do deseti cigaret z krabičky mi kamarádi přisypali do tabáku acylpyrin. Jaká je pravděpodobnost, že vykouřím osm normálních a dvě s acylpyrinem, když dohromady vykouřím 10 cigaret?

Můj výběr cigaret, jehož pravděpodobnost počítáme, bude obsahovat osm normálních cigaret z 10ti a dvojici acylpyrinem ochucených cigaret z 10ti. Ke každé možné skupině osmi cigaret můžeme nakombinovat libovolnou dvojici, proto tyto čísla násobíme. Výsledkem bude počet možností, kterými se dá vybrat 10 cigaret z 20ti tak, aby 8 bylo normálních a 2 s acylpyrinem. Celkový počet možností, jak se dá vybrat 10 cigaret z 20ti zjistíme taktéž kombinačním číslem.

$$\frac{\binom{10}{8}.\binom{10}{2}}{\binom{20}{10}} = \frac{45.45}{184756} = 0,011$$

2.2 Náhodná veličina

Výsledky pokusů (činnosti, při které můžeme dostat vícero výsledků) jsou často čísla, například výsledkem hodu kostkou je číslo, výsledný počet poruch za směnu je též číslo, atd. Toto číslo můžeme nazvat náhodnou veličinou. Tato veličina může nabývat různých hodnot, v případě kostky naše **náhodná veličina X** (tak se značí) může nabývat hodnoty 1,2,3,4,5 a 6. Počet poruch by se mohl pohybovat od 0 až do nekonečna a náhodná veličina by mohla teoreticky nabýt jakékoli hodnoty v tomto rozsahu.
Každá konkrétní hodnota náhodné veličiny X má i svoji pravděpodobnost, tedy pokud na kostce může nabývat X hodnoty od 1 do 6, tak umíme určit pravděpodobnost pro X=1, X=2 atd. Pravděpodobnost že padne jedno ze šesti čísel je 1/6, tedy i pravděpodobnost X=1 bude 1/6, zapisujeme P(X=1)=1/6. To samé bude platit i pro ostatní hodnoty X, protože každé číslo na kostce má stejnou pravděpodobnost. Právě jsme přiřadili každé možné hodnotě X pravděpodobnost, definovali jsme takzvanou **pravděpodobnostní funkci**. Vypadá takto:

P(X=1)=1/6	P(X=4)=1/6
P(X=2)=1/6	P(X=5)=1/6
P(X=3)=1/6	P(X=6)=1/6

Z této funkce můžeme lehce odvodit druhou důležitou funkci, a to **distribuční**, která nám udává pravděpodobnost, že náhodná veličina X nabude hodnoty **menší nebo rovné** nějakému číslu. Pro naši kostku by vypadala takto:

F(x) =0 pro x<1
=1/6 pro 1≤x<2
=2/6 pro 2≤x<3
=3/6 pro 3≤x<4
=4/6 pro 4≤x<5
=5/6 pro 5≤x<6
=1 pro x≥6

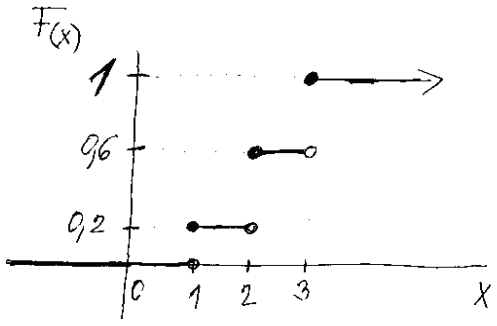
Distribuční funkci definujeme jako $F(x) = P(X \leq x)$, tedy funkční hodnota ve zvoleném bodě x se rovná pravděpodobnosti, že náhodná veličina X nabude hodnoty menší nebo rovné jako námi zvolený bod. Lidsky řečeno: Když si zvolíme v případě naší kostky za X trojku, zapisujeme $F(3)$, tak distribuční funkce nám řekne, jaká je pravděpodobnost, že na kostce padne číslo menší nebo rovno třem, zapisujeme $P(X \leq 3)$. Jak vidíme z tabulky výše, $P(X \leq 3) = \frac{3}{6}$.

9. Veličina X nabývá hodnot 1,2 nebo 3. Známé jsou pravděpodobnosti P(1)=0,2, P(2)=0,5. Určete chybějící pravděpodobnost P(3). Dále vypočítejte a interpretujte hodnotu distribuční funkce v bodě 2.

Jak víme, náhodná veličina nabývá pouze hodnoty 1,2 a 3, a zároveň máme zadané pravděpodobnosti, že *„padne*" jednička (0,2) i dvojka (0,5). Pravděpodobnost trojky bude tedy 0,3, protože základní vlastností pravděpodobnosti je, že součet pravděpodobností všech možných variant je 100%. Distribuční funkce v bodě 2 říká, jaká je pravděpodobnost, že výsledek náhodné veličiny bude číslo menší nebo rovné dvěma. Je to tedy součet pravděpodobnosti, že *"padne"* jednička a dvojka = 0,7.

10. 20% rodin má v domě jednu místnost, 40% jich má dvě a 40% má tří. Pro veličinu počet místností načrtněte graf distribuční funkce. Jakou má hodnotu v bodě 2? Co tato hodnota znamená?

x	$F(x)$
$x<1$	0
$1\leq x<2$	$0,2$
$2\leq x<3$	$0,6$
$3\leq x$	1



Při tvorbě grafu distribuční funkce nespojité veličiny (takové, které nenabývá hodnoty v každém bodu) dáváme pořád uzavřené intervaly na levou stranu, a to z toho důvodu, že distribuční funkce udává pravděpodobnost, že X nabude hodnoty menší nebo rovné než námi zvolené x. Když se podívám na graf a zeptám se, jaká je pravděpodobnost, že má rodina dvě nebo méně pokojů (hodnota distribuční funkce v bodě 2), vidím, že je to 0,6. Pro hodnoty x<1 je pravděpodobnost nulová, neboť nikdo nemá méně než 1 pokoj.

2.3 Rozdělení náhodné veličiny

Po hodu kostkou nám padne nějaké číslo. Umíme vypočítat, s jakou pravděpodobností dané číslo padne. Podobně, i když máme deset aut, přičemž jedno je porouchané, umíme vypočítat, s jakou pravděpodobností si vybereme právě to s poruchou. Statistika však dokáže vypočítat o hodně složitější věci a nástrojem na řešení těchto složitějších problémů jsou rozdělení náhodné veličiny. **Rozdělení**, to je takový **všeobecný vzorec**, který popisuje chování náhodné veličiny za různých podmínek[1] a zautomatizuje počítání do takové míry, že si jen musíme vybrat to správné rozdělení vzhledem k charakteru našeho příkladu a doplnit do vzorce proměnné. Představme si to, jako kdybychom napsali vzorec na pravděpodobnostní funkci jakkoli velké hrací kostky. Do toho vzorce bychom potom už jen doplnili, kolik má kostka stran, kolikrát házeme, jaké číslo chceme a kolikrát chceme, aby padlo – a vzorec nám už jen vychrlí výsledek – naši hledanou pravděpodobnost. Jednoduše řečeno, rozdělení náhodných veličin jsou vzorce, které nakrmíme nějakými vstupními proměnnými, a ony nám dopočítají hodnotu výsledné pravděpodobnostní nebo distribuční funkce. Tedy, dají nám odpověď na otázku (v případě pravděpodobnostní funkce): "Jaká je pravděpodobnost, že X (např. počet poruch) se bude rovnat 3?" a nebo (v případě distribuční funkce): "Jaká je pravděpodobnost, že X bude menší nebo rovno deseti?" Existují různé typy příkladů a na **základě jejich specifik je možné použít různá rozdělení** – proto u každého rozdělení, které budu zmiňovat, bude uvedené, na jaký typ příkladu je možné jej použít.

Binomické rozdělení – nám dokáže vypočítat **pravděpodobnost, že se v sérii pokusů (n) bude vyskytovat jev, který má nějakou pravděpodobnost (p) právě X krát**. Je to rozdělení, které musíme nakrmit dvěma proměnnými, a to proměnnou ***n***, která značí počet nezávislých náhodných pokusů *(např. počet hodů kostkou)* a proměnnou ***p***, která říká, jaká je pravděpodobnost jevu, který sledujeme *(pravděpodobnost, že padne šestka)*. Velké X je naše náhodná veličina, jejíž pravděpodobnost chceme zjistit *(zadáme-li X=3, počítáme pravděpodobnost, že šestka padne právě třikrát)*. Ve vzorci pro výpočet pravděpodobnostní funkce tohoto rozdělení je i proměnná ***q*** - ta se však jenom dopočítává jako $1-p$ a je to tedy pravděpodobnost, že daný jev nenastane. Pro úplnost je tady vzorec:

$$P(X=x)=\binom{n}{x}p^x q^{n-x}$$

11. Jaká je pravděpodobnost, že v pěti hodech kostkou padne 6 nejvýše jednou a jaká je pravděpodobnost, že padne alespoň třikrát?

Takže typické, série pokusů, je jich n, pravděpodobnost, že padne konkrétní číslo na kostce, je stará známá 1/6. Je třeba si uvědomit, že zjišťujeme pravděpodobnost náhodné veličiny, kterou je "počet padnutí šestky v pěti hodech". Ta může nabývat hodnoty od 0 až do 5. Pokud chceme zjistit pravděpodobnost, že padne nejvýše jednou, tedy 0 nebo jedenkrát, bude se tato pravděpodobnost rovnat součtu $P(X=0)$ a $P(X=1)$. Obdobně pro "alespoň třikrát", tedy 3, 4 a 5 je to buď součet pravděpodobností $P(3)+P(4)+P(5)$ a nebo $1-P(0)-P(1)-P(2)$, protože kdybychom sčítali pravděpodobnosti od 0 až po 5, tak nám musí vyjít 100%, že jedna z nich nastane. Můžeme tedy od celku (100%=1) odečíst, co chceme, a vyjde nám pravděpodobnost, že nastane to, co jsme neodečetli.

[1] Chování náhodné veličiny za různých podmínek = jaké hodnoty nabude náhodná veličina (počet padnutí šestky na hrací kostce), při měnících se podmínkách (počet stran kostky, počet hodů, atp.)

a.) nejvýše jednou

$$P(X=x)=\binom{n}{x}p^x q^{n-x}$$

$$p=\frac{1}{6} \quad q=1-\frac{1}{6}=\frac{5}{6} \quad n=5 \quad X=0$$

$$P(X=0)=\binom{5}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^5=1.1.\left(\frac{5}{6}\right)^5=0,4019$$

$$P(X=1)=\binom{5}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^4=5.\frac{1}{6}.\left(\frac{5}{6}\right)^4=0,4019$$

$$P(X=0)+P(X=1)=0,4019+0,4019=\underline{0,804}$$

b.) alespoň třikrát

$$P(X=2)=\binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^4=10\cdot\frac{1}{36}\cdot\frac{125}{216}=0,1608$$

$$1-P(X=0)-P(X=1)-P(X=2)=1-0,804-0,1608=\underline{0,0352}$$

Užitečná rada: Pokud máte nějakou vědečtější kalkulačku, určitě na ní najdete funkce, které vám mohou výrazně ulehčit počítání. Kombinační čísla se řeší tak, že nejprve vložíte do kalkulačky horní číslo, potom stisknete tlačítko nCr (já ho mám jako druhou funkci na tlačítku dělení) a potom zadáte spodní číslo. Rovná se a výsledek je výsledkem kombinačního čísla. Dobré je také používat zlomky, tlačítko je většinou označené jako $a\ b/c$. Zadáte čitatel, stisknete tlačítko, zadáte jmenovatel a zlomek je hotový. Tímto tlačítkem taktéž přepínáte mezi zlomkovým zobrazením a klasickým desetinným. Pozor, pokud chcete zlomek umocnit, musíte jej nejprve dát do závorky.

12. Jaká je pravděpodobnost, že si z deseti tahů vytáhneme z balíčku alespoň jedenkrát eso? Po vytáhnutí vracím kartu pokaždé zpátky a promíchám.

Stačí mi samozřejmě spočítat pravděpodobnost, že si eso nevytáhnu ani jedenkrát.

$$P(X=x)=\binom{n}{x}p^x q^{n-x}$$

$$p=\frac{4}{52}=\frac{1}{13} \quad q=1-\frac{1}{13}=\frac{12}{13} \quad n=10 \quad X=0$$

$$P(X=0)=\binom{10}{0}\left(\frac{1}{13}\right)^0\left(\frac{12}{13}\right)^{10}=1.1.\left(\frac{12}{13}\right)^{10}=0,45$$

$$P(X>0)=1-P(X=0)=1-0,45=\underline{0,55}$$

Pravděpodobnost, že si vytáhnu alespoň jedenkrát eso, je 55%.

Poissonovo rozdělení – má stejnou oblast použití jako rozdělení binomické. Také do něj vstupují proměnné n (počet pokusů) a p (pravděpodobnost zkoumaného jeve), jenom jej používáme v

případech, kdy počet prvků (teda ***n***) je více než 30 a pravděpodobnost (***p***) je malá, prakticky stačí, že je menší než 0,1, tedy 10%. Při splnění těchto podmínek můžeme říct, že Poissonovo rozdělení **aproximuje** rozdělení binomické – což znamená, že výsledky pří použití tohoto vzorce jsou jen minimálně odlišné od výsledků za použití vzorce pro binomické rozdělení. Ve vzorci Poissonova rozdělení se počítá s parametrem λ - lambda, který se rovná $p \cdot n$. Ve vzorci figuruje i konstanta e - Eulerovo číslo (zaokrouhlené 2,71828). Pravděpodobnost X se vypočítá pomocí rovnice:

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad \lambda = p \cdot n$$

Druhou důležitou oblastí použití Poissonova rozdělení jsou Poissonovské proudy. Zpravidla jde o příklady, ve kterých řešíme **pravděpodobnost výskytu nějakého jevu** (např. počet výběrů z bankomatu) **za určitou časovou jednotku** (např. 10 minut). Lambda je v takovém případě takzvaný parametr proudu a udává počet výskytů jevu za určitou časovou jednotku. Ve vzorci se nachází ještě jeden další parametr a to je t, které udává velikost sledovaného intervalu *(ve kterém chceme zjistit pravděpodobné množství výskytů jevu, který sledujeme)* jako zlomek celé časové jednotky.

$$P(x) = \frac{(t\lambda)^x}{x!} e^{-t\lambda}$$

13. Výběry z bankomatu se řídí Poissonovým rozdělením a za hodinu si z něj vybere peníze průměrně 40 lidí. Jaká je pravděpodobnost, že v průběhu následujících 5ti minut si nikdo nic nevybere?

Informace, že za hodinu si vybere peníze 40 lidí, nám určuje hodnotu parametru lambda – počet jevů za časovou jednotku (časová jednotka = hodina). Interval pěti minut, ve kterém chceme určit pravděpodobnost, musíme zohlednit jako parametr t. Má to být zlomek celé časové jednotky, pokud časová jednotka je hodina, 60 minut, tak 5 minut je jednou dvanáctinou. Proto t = 1/12. Náhodná veličina X, kterou zkoumáme, je počet jevů za interval – zkoumáme pravděpodobnost, že se nezrealizuje žádný výběr, proto budeme za X dosazovat nulu.

$$\lambda = 40 \qquad t = \frac{1}{12} \qquad e = 2,71828$$

$$P(X=0) = \frac{\left(\frac{1}{12} \cdot 40\right)^0}{0!} \cdot e^{-\frac{1}{12} \cdot 40} = \frac{1}{1} \cdot e^{-3\frac{1}{3}} = 2,71828^{-3\frac{1}{3}} = \underline{0,0357}$$

Hypergeometrické rozdělení – **používáme při výběru bez vracení,** což znamená, že každý další výběr je ovlivněný tím, co se vytáhlo v předchozím tahu – používá se tedy pro závislé náhodné veličiny. Obzvlášť časté jsou příklady, ve kterých máme pomíchané dva druhy něčeho *(např. bílé a černé kuličky, shnilé a zdravé jablka, přičemž po vytáhnutí z pytlíku to nevracíme zpátky a taháme dál)*. Parametry tohoto rozdělení jsou: ***N*** – počet všech prvků, ***M*** – počet prvků s nějakou specifickou vlastností *(např. shnilé jablka)*, ***n*** – počet prvků, které taháme, a konečně naší záhadnou veličinou, jejíž pravděpodobnost hledáme je ***x*** a označuje počet prvků z těch, které jsme vytáhli, které mají tu specifickou vlastnost, tedy např. kolik z těch jablek je shnilých.

$$P(x) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$$

14. Máme 10 výrobků a 4 z nich jsou zmetky, taháme 4 bez vracení. Jaká je pravděpodobnost, že alespoň jeden z nich bude zmetek?

$$N = 10 \qquad M = 4 \qquad x = 0 \qquad n = 4$$

$$P(X \geq 1) = 1 - P(X = 0)$$

$$P(X = 0) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{4}{0} \cdot \binom{6}{4}}{\binom{10}{4}} = \frac{1 \cdot 15}{210} = 0,0714$$

$$P(X \geq 1) = 1 - 0,0714 = \underline{0,9286}$$

Stačí nám zjistit, jaká je pravděpodobnost, že bude 0 zmetků, to odečteme od 100 procent a máme výsledek. Zmetky dosazujeme za M, celkový počet výrobků je N, taháme z nich 4, což dosadíme za proměnnou n.

15. Spolubydlící šel na nákup do Hypernovy, je však zapomnětlivý typ, proto jsme mu vytvořili mnemotechnickou pomůcku, pomocí které určitě nakoupí všechno důležité. Jmenuje se to zákon 4-3-2-1 a spočívá v tom, že každý nákup by měl obsahovat 10 věcí: 4 piva, troje chipsy, 2krát salát a jedny cigarety. Navzdory tomuto propracovanému plánu se mu podařilo zapomenout koupit 5 věcí. Jaká je pravděpodobnost, že koupil alespoň jedno pivo?

$$N = 10 \qquad M = 4 \qquad n = 5 \qquad x = 0$$

$$P(X = 0) = \frac{\binom{4}{0} \cdot \binom{10-4}{5-0}}{\binom{10}{5}} = \frac{1 \cdot 6}{252} = 0,024$$

$$P(X \geq 1) = 1 - P(X = 0) = 1 - 0,024 = \underline{0,976}$$

Takže žádný strach, pivo snad bude.

16. V zásilce 20ti výrobků jsou 2 zmetky. Náhodně taháme 5 kusů. Jaká je pravděpodobnost, že vytáhneme jeden zmetek, pokud taháme s vracením, a jaká bez vracení?

a.) s opakováním

$$x = 1 \qquad n = 5 \qquad p = 20/2 = 0,1 \Rightarrow q = 0,9$$

$$P(1) = \binom{n}{x} \cdot p^x \cdot q^{n-x} = \binom{5}{1} \cdot 0,1^1 \cdot 0,9^4 = 0,328$$

b.) bez opakování

$x=1 \quad n=5 \quad N=20 \quad M=2$

$$P(X=1)=\frac{\binom{2}{1}\cdot\binom{20-2}{5-1}}{\binom{20}{5}}=\frac{2\cdot 3060}{15504}=0,3947$$

17. Pěstovatel nakoupil 40 sazenic jabloní. Špatným skladováním došlo k tomu, že 8 jich uschlo. Jaká je pravěpodobnost, že při náhodném výběru 20ti sazenic (bez vracení) budou:

- *všechny dobré*
- *uschlé?*

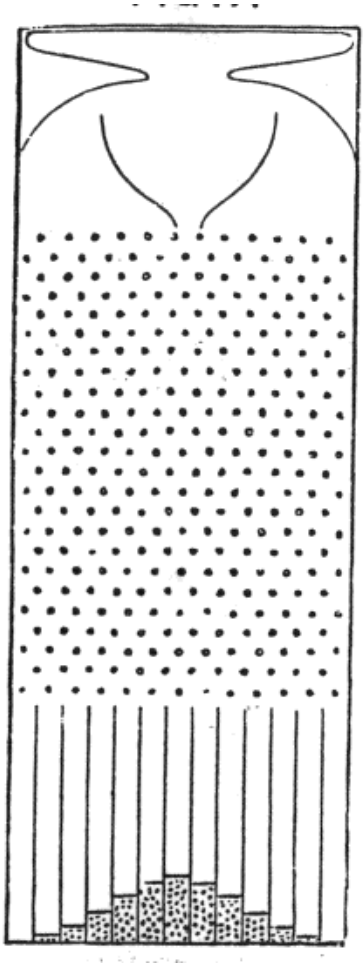
a.) $N=40 \qquad M=8 \qquad n=20 \qquad x=0$

$$P(x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}=\frac{\binom{8}{0}\binom{32}{20}}{\binom{40}{20}}=\frac{\frac{32\cdot 31\cdot 30\cdot\ldots\cdot 21}{12!}}{\frac{40\cdot 39\cdot 38\cdot\ldots\cdot 21}{20!}}=\frac{\frac{1}{1}}{\frac{40\cdot 39\cdot 38\cdot\ldots\cdot 33}{20\cdot 19\cdot 18\cdot 17\cdot 16\cdot 15\cdot 14\cdot 13}}=\frac{1}{610,5}=0,00164$$

b.) $N=40 \qquad M=8 \qquad n=20 \qquad x=4$

$$P(x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}=\frac{\binom{8}{4}\binom{32}{16}}{\binom{40}{20}}=\frac{70\cdot\frac{32\cdot 31\cdot 30\cdot\ldots\cdot 17}{16!}}{\frac{40\cdot 39\cdot 38\cdot\ldots\cdot 21}{20!}}=\frac{70\cdot\frac{20\cdot 19\cdot 18\cdot 17}{1}}{\frac{40\cdot 39\cdot 38\cdot\ldots\cdot 33}{20\cdot 19\cdot 18\cdot 17}}=\frac{8139600}{26666640}=0,305$$

Pravděpodobnost, že budou všechny dobré je 0,16%, a pravděpodobnost, že budou 4 uschlé je 30,5%.



Normální rozdělení – je nejdůležitějším pravděpodobnostním rozdělením a používá se hlavně jako **model pro rozdělení náhodných chyb měření**, které jsou způsobené množstvím malých, na sobě nezávislých náhodných jevů. Obecně normální rozdělení říká, že výsledky náhodných pokusů se budou pohybovat okolo průměrné hodnoty, přičemž to, jak moc budou okolo ní roztroušené, závisí na rozptylu náhodného jevu, který sledujeme. Proto má toto rozdělení dva parametry, a to μ - mí, které je **totožné s průměrem** a σ^2 - sigma na druhou, který je **totožný s rozptylem**.

Nejlepším příkladem ilustrujícím veličinu s náhodným rozdělením je model, který jsem viděl v muzeu vědy a techniky v Paříži. Nazývá se quincunx nebo **Galton Box**, podle svého zhotovitele Sira Francise Galtona. Jednalo se o jednoduché zařízení, které shora pouštělo asi sto úplně stejných kuliček. Tyto kuličky padaly na malé ocelové trubičky a byly nastavené tak, aby spadly přesně na střed prostřední trubičky. Z ní spadli buď vpravo anebo vlevo. Následně dopadly na další trubičku, ze které zase spadli buď vlevo, nebo vpravo. Tak to pokračovalo, až dokud kulička neprošla všemi poschodími zařízení a nakonec dopadla do jednoho z chlívků na jeho dně. Situaci ilustruje obrázek vlevo. Pravděpodobnost říká, že tak, jako hody mincí by měly být rovnoměrně rozdělené – kolikrát padne panna, tolikrát by měl padat i orel – tak i kuličky v zařízení by se měli odrážet jednou doprava, jednou doleva ve stejném poměru. Jak vidíme, není to úplně tak. Většina kuliček sice poměr víceméně dodržela a skončila nakonec ve středním chlívku nebo velmi blízko středu, jiné se častěji odrazily na pravou stranu, ostatní zase na levou stranu a skončily v krajnějších chlívcích. Jak vidíme, jejich rozmístění vymodelovalo klasickou křivku normálního rozdělení. Po skončení jednoho cyklu se kuličky vrátily zpět nahoru a pokus se opakoval. Kuličky potom vytvořily v chlívcích jinou křivku, která byla trochu špičatější, a na levé straně nebylo tolik koulí jako na pravé, ale je jasné, že pohyb těchto kuliček skrz zařízení se řídí normálním rozdělením. Odchylky v uspořádání kuliček po ukončení cyklu na dně stroje jsou náhodné nebo zapříčiněné množstvím nepatrných jevů, ale stále je kuliček nejvíc ve středu – okolo střední hodnoty – a v jeho okolí jsou rozptýlené vzhledem k rozptylu, který tento jev má.

Pojďme nyní k praxi. To, co většinou potřebujeme v příkladech na normální rozdělení vypočítat, je jeho distribuční funkce. Takže například pravděpodobnost, že rozměr součástky bude menší nebo stejný jako je norma, pravděpodobnost, že počet zápalek v krabičce je menší nebo roven 40ti, a tak podobně. V předešlých rozděleních jsme měli vzorce, do kterých jsme dosadili potřebné proměnné a hodnotu X, a dostali jsme pravděpodobnost. Takový vzorec by ale byl u normálního rozdělení příliš složitý a počítalo by se nám s ním nepohodlně a těžko. Proto si hodnotu náhodné veličiny X musíme upravit do takzvaného normovaného tvaru – na **normovanou veličinu U**. Potom bychom měli teoreticky s touto hodnotou U počítat dál a dospět k hodnotě pravděpodobnosti, kterou hledáme. Máme to ale ulehčeno tím, že pro konkrétní hodnoty U jsou uvedené v tabulkách konkrétní pravděpodobnosti, takže nám stačí vypočítat U, podívat se do tabulek a máme výsledek. Vzorec normované veličiny U obsahuje průměr μ, **odmocninu** z rozptylu $=\sigma$, a X, což je nějaká hodnota náhodné veličiny.

$$U = \frac{x - \mu}{\sigma}$$

Poslední stručná, důležitá a výstižná věta: Pokud chci zjistit distribuční funkci normálního rozdělení, tedy pravděpodobnost, že náhodná veličina bude nabývat hodnoty menší nebo stejné jako mnou zadané X, vypočítáme normovanou veličinu U, potom se podíváme do tabulek, kde na základě toho, kolik nám ta veličina U vyšla, zjistíme příslušnou hodnotu distribuční funkce, značíme $\Phi(U)$. Tato hodnota je hledanou pravděpodobností. Distribuční funkce udává pravděpodobnost, že X bude menší nebo rovno nějakému číslu, takže jestli se ptají na pravděpodobnost, že X bude větší než nějaké číslo, logicky musíme tu hodnotu distribuční funkce příslušící našemu U (výslednou pravděpodobnost) odečíst od 1: P(X>x)=(1-$\Phi(U)$).

18. Hmotnost vyráběných součástek je normálně rozdělená veličina se střední hodnotou 110 gramů a rozptylem 100. S jakou pravděpodobností bude hmotnost součástky menší nebo rovná 115ti gramům?

$$\underline{N(110;100) \quad \sigma = \sqrt{100} = 10 \qquad x = 115}$$

$$U = \frac{x-\mu}{\sigma} = \frac{115-110}{10} = 0,5$$

$$P(X \le 115) = \Phi U = \Phi(0,5) = 0,691 = \underline{69,1\%}$$

Takže – ptají se na pravděpodobnost, že bude hmotnost menší nebo rovná 115ti gramům, z toho vyplývá, že se ptají na distribuční funkci, a že za X dosadíme 115. Rozptyl je 100, za sigmu dosadíme jeho odmocninu, tedy 10, za μ dosadíme 110, tedy střední hodnotu. Vypočítáme U, a potom se už jenom podíváme do tabulek (konkrétně tabulka „Distribuční funkce normovaného normálního rozdělení") a příslušná hodnota pravděpodobnosti (resp. Hodnota F(u)) k našemu vypočítanému U je výsledek.

19. V testech inteligence je průměrný výsledek 100 bodů se směrodatnou odchylkou 15 bodů. Kolik procent lidí dosáhne více než 105ti bodů a kolik procent lidí dosáhne maximálně 90ti bodů? V jakém intervalu symetrickém okolo střední hodnoty se bude nacházet 50% lidí?

Pokud chceme zjistit procento pro více než 105 bodů, vypočítáme procento pro 105 a méně a potom jej odečteme od jedničky. Maximálně 90 bodů je distribuční funkce v bodě 90 – tedy procento lidí s počtem bodů menším nebo rovným 90ti.

$$\underline{N(100;225) \quad \sigma = 15 \qquad x > 105}$$

$$U = \frac{x-\mu}{\sigma} = \frac{105-100}{15} = 0,333$$

$$P(X \le 105) = \Phi U = \Phi(0,333) = 0,62930$$

$$P(X > 105) = 1 - 0,62930 = 0,3707$$

$$\underline{N(100;225) \quad \sigma = 15 \qquad x \le 90}$$

$$U = \frac{x-\mu}{\sigma} = \frac{90-100}{15} = -0,666$$

$$\Phi(-U) = 1 - \Phi(U)$$

$$P(X \le 90) = \Phi(-0,666) = 1 - \Phi(0,666) =$$

$$= 1 - 0,74537 = 0,25463$$

V případě, že hodnota normované veličiny U je záporná (záporné hodnoty nejsou v tabulkách), vypočítáme hledanou pravděpodobnost jako 1-(pravděpodobnost kladné hodnoty U). Na otázku, v jakém intervalu se bude nacházet 50% lidí, odpovíme, že v intervalu (85;115). To jsme vypočítali jako

střední hodnota$\pm$ směrodatná odchylka. Vyplývá to z definice směrodatné odchylky, která říká, že směrodatná odchylka je hodnota, o kterou se v obou směrech neodchyluje od průměru více než 50% hodnot.

20. Nespojitá celočíselná náhodná veličina X má normální rozdělení se střední hodnotou 7 a rozptylem 4. Určete pravděpodobnost, že tato náhodná veličina nabude hodnot:

a. *maximálně 6*
b. *aspoň 5*
c. *z intervalu (5,9)*

$N(7;4) \qquad \sigma = \sqrt{4} = 2$

$$P(X \le 6) = \Phi U = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{6-7}{2}\right) = \Phi(-0,5) = 1-\Phi(0,5) = 1-0,69 = 0,31$$

$$P(X \ge 5) = 1-P(X \le 4) = 1-\Phi\left(\frac{4-7}{2}\right) = 1-\Phi\left(-\frac{3}{2}\right) = 1-(1-\Phi(1,5)) = 1-(1-0,933) = 0,933$$

$$P(5 \le X \le 9) = \Phi\left(\frac{9-7}{2}\right) - \Phi\left(\frac{5-7}{2}\right) = \Phi(1) - (1-\Phi(1)) = 0,682$$

V prvním případě je to jen jednoduchá distribuční funkce. Ve druhém případě vyjádříme pravděpodobnost, že X nabude hodnoty menší než 4 a tu odečteme od 1 (protože veličina je celočíselná). Ve třetím případě musíme od pravděpodobnosti, že X bude menší než 9 odečíst ještě pravděpodobnost, že bude menší než 5, a tak dostaneme pravděpodobnost, že se bude nacházet v zadaném intervalu.

21. Hmotnost výrobku je vyhovující, pokud je v rozmezí 68 - 69 gramů. Za standardních podmínek má hmotnost přibližně normální rozdělení se střední hodnotou μ =68,3 gramů a směrodatnou odchylkou v předepsaných mezích. Jaká je pravděpodobnost, že hmotnost výrobku bude vyhovující?

$N(68,3;0,09) \qquad \sigma = 0,3$

$$P(68 \le X \le 69) = \Phi\left(\frac{69-68,3}{0,3}\right) - \Phi\left(\frac{68-68,3}{0,3}\right) = \Phi(2,33) - \Phi(-1) =$$

$$= 0,99 - (1-\Phi(1)) = 0,99 - 0,159 = 0,831 = 83,1\%$$

Je-li směrodatná odchylka v předepsaných mezích, může být maximálně 0,3, protože jinak by ve směru dolů překročila limit (68 < střední hodnota ± směrodatná odchylka < 69). Hledáme pravděpodobnost, že hmotnost bude v mezích – mezi 68 a 69. Vypočítáme pravděpodobnost, že bude menší než 69, ale od ní musíme ještě odečíst pravděpodobnost, že bude menší než 68.

3 Bodový a intervalový odhad

Pokud statisticky zjišťujeme nějaký jev, často nastává situace, že rozsah souboru je tak velký, že je velmi obtížné zjistit skutečný stav. Tedy, pokud zjišťujeme preference politických stran, je samozřejmé, že se nemůžeme ptát každého občana, koho bude volit. Obdobně i u testování součástek nemůžeme otestovat všechny, ale jen nějaký vzorek, tzv. výběrový soubor. Příkladů ze života si i sami domyslíte spoustu. Pro nás je důležité to, že **základní soubor** *(celá populace, všechny součástky)* má svoje statistiky jako jsou rozptyl a průměrná hodnota. Stejně má nějaký průměr, resp. Rozptyl i **výběrový soubory (tedy ten vzorek)** a my ve valné většině příkladů chceme na základě dat, které máme z výběrového souboru, určit průměr nebo rozptyl základního souboru, přesněji řečeno, určit interval, ve kterém se tyto statistiky nachází.

3.1 Teoretický úvod

Ještě jednou – jde nám o to, **zjistit buď rozptyl, nebo střední hodnotu základního souboru**, pokud **známe hodnoty** nějakého **výběru** n prvků. Celý postup výpočtu spočívá v dosazování do vzorců. Abychom ale uměli vzorce taky použít, je třeba nejprve pochopit dvě základní věci:

Věc 1: Pokud zjišťujeme rozptyl nebo střední hodnotu základního souboru (ZS) na základě nějakého výběru z něj, konečný výsledek udáváme v intervalu, protože není možné určit přesnou hodnotu *(pokud se např. zeptáme tisíce občanů ČR na jejich výšku a průměr z toho, co nám uvedou, vyjde 170, nemůžeme jednoduše prohlásit, že průměr výšky všech obyvatel ČR je 170)*. Nemůžeme ale říct, že např. „průměrná výška obyvatel ČR (střední hodnota ZS) se bude na 100% nacházet v intervalu 150 až 200". Výsledkem je interval, ve kterém se bude nacházet hledaná proměnná, a i to pouze s určitou pravděpodobností. V praxi se však často střetáváme s problémem, že pokud určujeme interval se stoprocentní pravděpodobností, že se v něm bude nacházet zjišťovaná neznámá, je tento interval tak široký, že je nám to na nic. Takže pokud děláme průzkum preferencí na vzorku 1000 lidí a jako výsledek uvedeme, že preference Demokratické strany u celé populace jsou na 100% v intervalu 10 až 20 procent, je to nepoužitelné. Proto se tyto intervaly uvádí na přesnostech nižších než 100%, a to obvykle s 95% přesností, která nám už poskytuje užší interval při málo změněné věrohodnosti. Přesnost, se kterou udáváme výsledek, nazýváme **konfidenční interval**, a značíme jako $1-\alpha$, přičemž α znamená vlastně možnou chybu odhadu. V případě $1-\alpha = 0,95$ je možná chyba 5%. Všechny vzorce, které se budou dále používat, budou cca ve tvaru:

$$P(X < zjišťovaná\ hodnota < Y) = 1-\alpha$$

Neboli: S pravděpodobností rovnou 1-α se bude zjišťovaná hodnota *(základního souboru = průměr nebo rozptyl)* nacházet v intervalu (X;Y).

Věc 2: Co je to „X" a „Y", které se píše ve vzorci o řádek výše?

$$P\left(\overline{x} - u_{1-(\alpha/2)}\frac{\delta}{\sqrt{n}} < \mu < \overline{x} + u_{1-(\alpha/2)}\frac{\delta}{\sqrt{n}}\right) = 1-\alpha$$

Levou a pravou hranici intervalu, ve kterém se bude nacházet zjišťovaná hodnota základního souboru (v případě střední hodnoty normálního rozdělení značíme μ), tvoří stále korespondující hodnota výběru ze základního souboru *(v tomto případě $\overline{x}$ - tedy pokud chceme znát střední hodnotu ZS, používá se tam střední hodnota výběru)*, od které je na levé straně odečtená a na pravé straně zase

přičtená jakási **chyba** odhadu, ve které je vždy zakomponovaný **kvantil nějakého rozdělení**, o kterém nemusíme vědět nic, kromě toho, že ho najdeme v tabulkách, kde jsou jeho hodnoty uspořádané podle pravděpodobnosti, se kterou interval určujeme, podle 1-α.

Shrnutí: Máme nějaký statistický soubor, který je velký, proto z něj vybereme několik exemplářů. Poznačíme si, kolik jsme jich vybrali, a zjistíme průměr a rozptyl. My ale chceme zanalyzovat průměr základního souboru. Takže pomocí toho, že jsme naměřili hodnoty té vybrané skupiny, dosadíme naměřená čísla a pár čísel, která najdeme v tabulkách, do vzorce, a ten nám vypočítá, v jakém intervalu se nachází námi hledaná hodnota základního souboru. Tento výsledek bude přesný podle toho, jaký zvolíme konfidenční interval. Čím chceme přesnost větší, tím větší bude interval, ve kterém se bude nacházet výsledek. Proto většinou volíme přesnost 95%, což je kompromis mezi šířkou intervalu, ve kterém se nachází výsledek, a přesností (pravděpodobností, že je ten výsledek správný).

3.2 Zjišťování střední hodnoty

V případě, že **zjišťujeme střední hodnotu** základního souboru, mohou nastat dvě situace. Buď známe rozptyl základního souboru (což je obvyklé spíš u poměrně nereálných příkladů), nebo rozptyl základního souboru neznáme a budeme muset použít rozptyl výběrového souboru, který může být zadaný, nebo jej budeme muset vypočítat podle vzorce:

$$s_x^{'} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Přeloženo: Od každé hodnoty výběru odečtu průměr celého výběru a výsledek umocním na druhou. Takovým způsobem to udělám se všemi hodnotami výběru a všechny výsledky umocněné na druhou sečtu. Potom to vydělím počtem hodnot zmenšeným o 1 a nakonec ještě odmocním.

Zjišťujeme střední hodnotu, známe rozptyl, používáme vzorec:

$$P(\overline{x} - u_{1-(\alpha/2)}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + u_{1-(\alpha/2)}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

1. Zjišťujeme průměrnou mzdu všech absolventek zdravotnické školy, přičemž pomocí předešlého zkoumání víme, že její rozptyl je 990 025. Vybrali jsme si náhodně 25 absolventek, u kterých jsme zjistili průměrnou mzdu 12 494 Kč. Sestrojte interval průměrné mzdy absolventek s přesností 95%.

Celá naše práce prakticky spočívá v dosazování do vzorce. Za $\overline{x}$ dosadíme průměrnou mzdu = 12 494. Za u musíme dosadit hodnotu příslušného kvantilu, kterou najdeme v tabulkách *(Kvantily normovaného normálního rozdělení)*. Nejprve musíme ale vědět, na jaké pravděpodobnosti počítáme. V zadání chtějí přesnost 95%, a to znamená, že $1-\alpha$ se bude rovnat 0,95 a $\alpha = 0,05$. Ve vzorci se ale píše, že dosazujeme kvantil $u_{1-(\alpha/2)}$, proto musíme spočítat kolik je $1-(\alpha/2)$. Je to 0,975 a v tabulkách najdeme hodnotu kvantilu příslušící pravděpodobnosti 0,975 a tou je 1,96. Tuto hodnotu dosadíme za výraz $u_{1-(\alpha/2)}$. Za sigmu dosadíme směrodatnou odchylku, což je odmocněný rozptyl, takže 995. A nakonec ještě dosadíme za n 25. Spočítáme čísla a vidíme, v jakých intervalech se bude μ základního souboru nacházet – to je náš výsledek.

$n = 25 \qquad \delta^2 = 990\,025 \qquad \overline{x} = 12\,494 \qquad 1-\alpha = 0,95$

$\sqrt{\delta^2} = \delta = 995$

$$P(\overline{x} - u_{1-\alpha/2} \cdot \frac{\delta}{\sqrt{n}} < \mu < \overline{x} + u_{1-\alpha/2} \cdot \frac{\delta}{\sqrt{n}}) = 1-\alpha$$

$$u_{1-\alpha/2} = u_{1-0,05/2} = u_{0,975} = 1,96$$

$$P(12\,494 - 1,96 \cdot \frac{996}{\sqrt{25}} < \mu < 12\,494 + 1,96 \cdot \frac{995}{\sqrt{25}}) = 0,95$$

$$P(12104 < \mu < 12\,884) = 0,95$$

S pravděpodobností 95% bude průměrná mzda v intervalu (12104;12884).

Zjišťujeme střední hodnotu, neznáme rozptyl

Příklady na výpočet parametrů základního souboru bez toho abychom znali jeho rozptyl jsou o hodně reálnější, jelikož v realitě opravdu většinou rozptyl základního souboru není známý. Postup výpočtu je však podobný jako když rozptyl známe až na dvě malé změny:

Když nemáme rozptyl, musíme na jeho místo dosadit něco jiné. Nazývá se to výběrový rozptyl. Označuje se s_x' a je to vlastně rozptyl našeho výběru. Jsou dvě možnosti – buď si ho musíme vypočítat ze zadaných hodnot podle vzorce:

$$s_x' = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

což znamená, že v případě mezd sestřiček z příkladu výše bychom museli znát mzdu každé z nich a potom bychom to mohli spočítat podobně jako rozptyl v začátcích. Druhá možnost je, že už bude zadaný a basta. Namísto kvantilu normovaného normálního rozdělení se používá **kvantil t Studentova rozdělení**. Pro nás to znamená, že namísto toho, abychom se dívali do tabulek na kvantily normálního rozdělení, najdeme si tabulku kvantilů Studentova rozdělení. Jediné, na co musíme dát pozor je, že Studentovo rozdělení má n-1 stupně volnosti, což v překladu znamená, že když je rozsah souboru n, dívám se do n-1 řádku v tabulkách *(pokud mám dvacet údajů, kvantil hledám v devatenáctém řádku).*

2. Známky z devíti náhodně vybraných předmětů, které jsem vystudoval během bakaláře, jsou: 2, 1, 3, 4, 3, 3, 2, 2, 3. Celkový počet mých vystudovaných předmětů je 41. Určete s pravděpodobností 95%, v jakém intervalu se bude nacházet můj průměr.

Nejdříve vypočítáme výběrový rozptyl, na ten ovšem potřebujeme průměr výběru:

$$\overline{x} = (2+1+3+4+3+3+2+2+3)/9 = 23/9 = 2,5\overline{5}$$

$$s_x^{'} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(1-2,5\bar{5})^2 + 3\cdot(2-2,5\bar{5})^2 + 4\cdot(3-2,5\bar{5})^2 + (4-2,5\bar{5})^2}{8}} = \sqrt{\frac{6,223}{8}} = 0,88197$$

Nyní zjistíme kvantil rozdělení t, v tabulkách jej najdeme v n-1 řádku, tedy osmém:

$$t_{1-\alpha/2} = t_{0,975} = 2,306$$

A nakonec vše dosadíme do vzorce:

$$P(\bar{x} - t_{1-(\alpha/2)}\frac{s_x^{'}}{\sqrt{n}} < \mu < \bar{x} + t_{1-(\alpha/2)}\frac{s_x^{'}}{\sqrt{n}}) = 1-\alpha$$

$$P(2,5\bar{5} - 2,306\cdot\frac{0,88197}{\sqrt{9}} < \mu < 2,5\bar{5} + 2,306\cdot\frac{0,88197}{\sqrt{9}}) = 0,95$$

$$P(1,88 < \mu < 3,23) = 0,95$$

Jak vidíme, výsledek je dost nepoužitelný, jeho interval je příliš široký. Zkusme proto snížit přesnost na úroveň $1-\alpha = 0,8$, tedy na 80%. Změní se nám v podstatě jenom kvantil t, který namísto hodnoty 2,306 bude mít hodnotu 1,397:

$$1-\alpha = 0,8 \Rightarrow \alpha = 0,2$$

$$t_{1-\alpha/2} = t_{1-0,2/2} = t_{0,9} = 1,397$$

$$P(2,5\bar{5} - 1,397\cdot\frac{0,88197}{\sqrt{9}} < \mu < 2,5\bar{5} + 1,397\cdot\frac{0,88197}{\sqrt{9}}) = 0,8$$

$$P(2,145 < \mu < 2,966) = 0,8$$

Interval se nám díky menší přesnosti výrazně zúžil. Můj skutečný průměr je 2,34.

Zjišťujeme střední hodnotu, všeobecné rozdělení, neznáme rozptyl, výběr n>30

V případě, že neznáme rozptyl, a velikost výběru je větší než 30, můžeme jakékoli rozdělení, kterým se daná náhodná veličina řídí, nahradit normovaným normálním rozdělením. Proto **v případě, že neznáme rozptyl ZS a n>30, používáme vzorec**:

$$P(\bar{x} - u_{1-\alpha/2}\cdot\frac{s_x^{'}}{\sqrt{n}} < E(X) < \bar{x} + u_{1-\alpha/2}\cdot\frac{s_x^{'}}{\sqrt{n}}) = 1-\alpha$$

ve kterém je místo μ (což je střední hodnota normálního rozdělení – neboli jsme předpokládali, že ZS bude mít normální rozdělení) uvedené E (X) jako všeobecné vyjádření střední hodnoty.

3. Na základě uvedené tabulky zaznamenávající pro 500 sledovaných rodin počet dětí a počet místnosti sestrojte 95% oboustranný interval spolehlivosti pro průměrný počet dětí.

Počet dětí v rodině	1	1	2	2	2	3	3
Počet místností	1	2	1	2	3	2	3
Četnost rodin (%)	10	10	10	20	10	20	20

$n = 500 \qquad 1-\alpha = 0,95 \qquad \overline{x} = ? \qquad s_x^{'} = ? \qquad \mu = ?$

$$\overline{x} = (0,2 \cdot 1) + (0,4 \cdot 2) + (0,4 \cdot 3) = 2,2$$

$$s_x^{'} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{100 \cdot (1-2,2)^2 + 200 \cdot (2-2,2)^2 + 200 \cdot (3-2,2)^2}{499}} = \sqrt{\frac{144+8+128}{499}} = 0,75$$

$$P(\overline{x} - u_{1-\alpha/2} \cdot \frac{s_x^{'}}{\sqrt{n}} < \mu < \overline{x} + u_{1-\alpha/2} \cdot \frac{s_x^{'}}{\sqrt{n}}) = 1-\alpha$$

$$P(2,2 - 1,96 \cdot \frac{0,75}{\sqrt{499}} < \mu < 2,2 + 1,96 \cdot \frac{0,75}{\sqrt{499}}) = 0,95$$

$$P(2,134 < \mu < 2,266) = 95\%$$

Nejprve si spočítáme střední hodnotu výběru *(výběr je těch 500 rodin = n)*, která znamená, kolik „dítěte připadá na rodinu". Dále spočítáme výběrový rozptyl jak je to demonstrované v řešení. V tabulkách najdeme kvantil pro pravděpodobnost 0,975. Dosadíme do vzorce a hotovo.

3.3 Odhad relativní četnosti základního souboru

Relativní četnost můžeme odhadovat, pokud je náhodná **veličina rozdělená alternativně**. Není to žádná věda, prakticky to znamená, že může nabývat pouze dvou stavů, tedy buď je jablko dobré, nebo je zkažené, buď se narodí kluk, nebo holka, atd. Zároveň je daná pravděpodobnost, s jakou jedna či druhá možnost nastanou *(je jasné, že „pravděpodobnost že nastane jeden jev" + „pravděpodobnost že nastane druhý" = 1, protože jeden z nich určitě nastane)*. Odhad relativní četnosti znamená, že v zadání je řečeno, jaká část výběru má nějakou vlastnost *(např. je zkažená)* a my z toho máme vypočítat interval, kolik procent základního souboru bude mít tuto vlastnost, samozřejmě s nějakou pravděpodobností, většinou 95%. V příkladech tedy zase jde jen o to správně identifikovat, že je to právě tento typ příkladu – použít správný vzorec a správně dosadit.

4. *Průzkumem se zjistilo, že 90 z 800 smrků je napadených kůrovcem. Sestrojte 95% interval pro podíl napadených smrků v celém lese.*

$90 \text{ z } 800 \text{ je napadených } \Rightarrow \; p = \frac{90}{800} = 0,1125 \qquad 1-\alpha = 0,95$

$$P(p - u_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} < \pi < p + u_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}) = 0,95$$

$$P(0,1125 - 1,96 \cdot \sqrt{\frac{0,1}{800}} < \pi < 0,1125 + 1,96 \cdot \sqrt{\frac{0,1}{800}}) = 0,95$$

$$P(0,0906 < \pi < 0,1344) = 0,95$$

Podíl napadených smrků v lese bude v intervalu (9,06% ; 13,44%).

Nejprve si musíme ujasnit pravděpodobnost, jsou dvě možnosti – buď je strom napadený anebo ne. Pravděpodobnost, že je napadený vyplývá z našeho výběru takže 90/800=0,1125. Teď se podívejme na vzorec:

$$P\left(p-u_{1-\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}<\pi<p+u_{1-\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}\right)=1-\alpha$$

Jak vidíme, dosazujeme jen „p", což je pravděpodobnost, že je strom napadený, potom příslušný kvantil „u", který najdeme v tabulkách a „n", tedy počet členů výběru. Uprostřed vzorce je proto π, protože tak se označuje střední hodnota alternativního rozdělení. Je také důležité si uvědomit, že pokud by se ptali v zadání na podíl **ne**napadených stromů v lese, museli bychom vypočítat pravděpodobnost, že je strom nenapadený a tu potom dosazovat za p.

5. *Mezi 75 kontrolovanými výrobky mělo 63 vyhovující jakost. Sestrojte 95% interval spolehlivosti pro podíl vyhovujících výrobků.*

Ze 75 výrobku je 63 vyhovujících $\Rightarrow$ pravděpodobnost vyhovujícího = $p=\frac{63}{75}=0,84$

$1-\alpha=0,95\Rightarrow u_{0,975}=1,96$

$$P\left(p-u_{1-\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}<\pi<p+u_{1-\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}\right)=1-\alpha$$

$$P\left(0,84-1,96\sqrt{\frac{0,84\cdot(1-0,84)}{75}}<\pi<0,84-1,96\sqrt{\frac{0,84\cdot(1-0,84)}{75}}\right)=0,95$$

$$P(0,757<\pi<0,923)=0,95$$

Na pravděpodobnosti 95% očekáváme, že podíl vyhovujících součástek bude v intervalu (75,7% ; 92,3%)

6. *Z 1500 náhodně dotázaných dospělých obyvatel města by určitou stranu volilo 225. Odhadněte se spolehlivostí 0,95 počet potencionálních voličů této strany ve městě, ve kterém žije 200 000 dospělých obyvatel.*

Z 1500 by stranu volilo 225 $\Rightarrow$ $p=\frac{225}{1500}=0,15$

$1-\alpha=0,95\Rightarrow u_{0,975}=1,96$

$$P\left(p-u_{1-\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}<\pi<p+u_{1-\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}\right)=1-\alpha$$

$$P\left(0,15-1,96\sqrt{\frac{0,15\cdot(1-0,15)}{1500}}<\pi<0,15+1,96\sqrt{\frac{0,15\cdot(1-0,15)}{1500}}\right)=0,95$$

$$P\left(0,132<\pi<0,168\right)=0,95$$

$$0,132\cdot 200\,000=26\,400$$

$$0,168\cdot 200\,000=33\,600$$

Princip zůstává stále stejný, jenom na závěr vynásobíme počet obyvatel procenty, které nám vyšly. Počet voličů se bude pohybovat s pravděpodobností 95% v intervalu (26400 ; 33600).

4 Testování hypotéz

Hypotéza je předpoklad o něčem. Věta „Zítra odpoledne bude 22 stupňů Celsia" by se též dala považovat za hypotézu. Naší úlohou ve statistice je danou hypotézu ověřit a vyjádřit se o její pravdivosti (otestovat ji). Proto budeme mít zadanou *(anebo si musíme sami vymyslet podle zadání příkladu)* takzvanou **nulovou hypotézu**, značíme H_0. Touto nulovou hypotézou je hypotéza, kterou testujeme, neboli o které chceme rozhodnout, zda je pravdivá. Proti této hypotéze stavíme **alternativní hypotézu** H_1, která musí popírat původní hypotézu.

Například, pokud výrobce lentilek garantuje, že v balíčku jich je 40, bude to naše nulová hypotéza H_0. Proti ní musíme postavit nějakou jinou, která to popírá. Nejčastěji je to jednoduché popření H_0, tedy „V balíčku není 40 lentilek". Jelikož je to prakticky negace H_0 značíme „H_1=non H_0". Pokud však vezmeme v úvahu specifikum příkladu, to, že není špatné, jestliže je v balíčku lentilek víc, můžeme sestavit i jinou H_1, a to že „Lentilek je v balíčku méně než 40". To by bylo něco teorie na úvod, ale pokud chceme tyto věci prakticky počítat, hlavní je, pochopit systém, který spočívá v práci se Vzorci a Tabulkami. Ve vzorcích najdeme tabulku pro testování hypotéz, velmi rozumně rozdělenou na tři sloupce, přičemž:

V prvním sloupci vždy najdeme hypotézu H_0, kterou chceme testovat. Protože jsme ve statistice, budeme testovat jen věci typu: „výrobce udává takovouto střední hodnotu blabla…" a „rozptyl mezd je takový a takový … a jaký je doopravdy?", a tedy to bude stále nějaký předpoklad o velikosti střední hodnoty nebo rozptylu a naše H_0 bude stále, že velikost **je** taková, jakou udává výrobce, že **je** taková, jak se předpokládá. Hned vedle si vybereme jednu z formulací H1, tedy buď úplně zamítneme hypotézu H_0, řekneme, že střední hodnota se **nebude** rovnat té, kterou předpokládáme ($\mu \neq \mu_0$), dá se psát i „non H_0" anebo jenom řekneme, že bude větší nebo menší.

Ve druhém sloupci máme **testové kritérium**. To je rovnice, ve které máme na levé straně nějakou neznámou, kterou potřebujeme vypočítat *(např. U nebo t)* a do pravé strany bychom měli dosazovat proměnné podle zadání, příp. z Tabulek. Až vypočítáme tuto proměnnou na levé straně, můžeme přikročit ke třetímu sloupci.

Ve třetím sloupci se nachází **kritický obor**. Označuje se W_a. **Patří-li proměnná**, kterou jsme vypočítali ve sloupci 2, **do tohoto kritického oboru**, potom **je platná hypotéza H_1.** Hypotéza H_0 je naopak neplatná. Pokud **proměnná nepatří do kritického oboru, tak hypotéza H_1 není platná** a pravděpodobně platí H_0. To však nemůžeme určit s jistotou, proto se takovému tvrzení vyhýbáme, většinou jenom konstatujeme, že H_1 není platná. Jak vidíme, kritické obory tam máme většinou 3 *(tři rovnice W_a = něco)*, ty jsou pod sebou uspořádané podle toho, kterou hypotézu H_1 jsme použili. Pokud jsme použili první H_1, díváme se na první kritický obor. Pokud druhou, tak se díváme na druhý a ostatní nás nezajímají. Kritický obor vždy porovnává vypočítanou proměnnou z druhého sloupce s nějakým kvantilem, který najdeme v tabulkách. Když nerovnost ve složených závorkách platí, značí to, že proměnná patří do kritického oboru= je neplatná hypotéza H_0.

Věřím, že absolutně jasné to bude z příkladů:

1. *Odchylky délky součástek od normy činí průměrně 16mm, po změně technologie bylo náhodně vybraných 50 a zjištěná odchylka byla 14,9mm se směrodatnou odchylkou 3,9mm. Otestuj na hladině významnosti 5% hypotézu, že technologie snižuje průměrnou velikost odchylek.*

$\mu_0 = 16\,mm \qquad n = 50 \qquad \bar{x} = 14,9\,mm \qquad s_x = 3,9\,mm$

$H_0 : E(X) = \mu_0 \qquad H_1 : E(X) < \mu_0 \Rightarrow W_a = \{U \le -u_{1-\alpha}\}$

$\alpha = 0,05 \Rightarrow 1-\alpha = 0,95 \qquad u_{0,95} = 1,645$

$U = \frac{\bar{x} - \mu_0}{s_x'} \cdot \sqrt{n} = \frac{14,9-16}{3,9} \cdot \sqrt{50} = -1,99$

$-1,99 \le -1,645 \Rightarrow platí$

$T \in W_a \Rightarrow zamítáme\ H_0;\ platí\ H_1$

Nejprve si musíme uvědomit, co vlastně počítáme. Pokud porovnáváme **střední hodnotu** odchylky součástek od normy před a po technologické změně, vybíráme si ze Vzorců tabulku, ve které jsou vzorce pro výpočet střední hodnoty. Ty jsou dvě a my si vybereme druhou (střední hodnota, všeobecné rozdělení, velký výběr), protože první se používá, když je výběr <30 a my jsme vybrali 50 součástek. Potom zformulujeme nulovou hypotézu. Možnost, jak je vidět z tabulky, máme jen jednu: $E(X) = \mu_0$, přičemž za μ_0 vždy považujeme původní střední hodnotu, v našem případě tu před změnou technologie. Takže naše nulová hypotéza říká, že nový průměr $E(X)$ *(po změně technologie)* těch odchylek součástek od normy je stejný jako byl původní. My máme zjistit, zda technologie **snižuje** průměrnou odchylku. Proto alternativní hypotézou H_1 bude, že průměrná velikost odchylek po změně bude **menší** než předtím. Tedy alternativní hypotéza H_1 = $E(X) < \mu_0$. Teď přejdeme do druhého sloupce a dosadíme do vzorce. $\bar{x}$ je vždy průměr po změně *(je to ale průměr výběru, zatímco $E(X)$, který zjišťujeme, je průměr celého souboru po změně technologie)*, μ_0 zase průměr původního souboru. Za výběrový rozptyl dosadíme směrodatnou odchylku a za n samozřejmě počet prvků výběru. Vypočítáme U *(výsledek testového kritéria – v tomto případě veličiny U – se často označuje všeobecně T)* a potom ještě najdeme v tabulkách správný kvantil pro pravděpodobnost 0,95 a po porovnání vidíme, že uvedená nerovnost *(druhá shora, protože jsme vybrali druhou H_0)* platí. Pokud nerovnost platí, jsme v kritickém oboru. To znamená, že platí hypotéza H_1. Střední hodnota odchylek po změně technologie je skutečně nižší než před změnou.

2. *Dlouhodobý průměr rozdaných letáků za jeden den u jednoho roznašeče je 1000 ks. Sledovali jsme konkrétního roznašeče letáků a naměřili jsme mu tyto denní hodnoty:*

Den	1	2	3	4	5	6	7
Počet	992	1024	890	970	1100	869	908

Ověřte, zda roznašeč letáků dosahuje alespoň dlouhodobého průměru na hladině $\alpha = 0,05$.

$\mu_0 = 1000 \qquad \alpha = 0,05 \qquad n = 7 \qquad \mu = ?$

$H_0 : \mu = \mu_0 \qquad t = \frac{\overline{x} - \mu_0}{s_x^{'}} \cdot \sqrt{n}$

$H_1 : \mu < \mu_0 \qquad \overline{x} = 6753 / 7 = 964,7$

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = (992 - 964,7)^2 + (1024 - 964,7)^2 + (890 - 964,7)^2 + (970 - 964,7)^2 + (1100 - 964,7)^2 +$$
$$+(869 - 964,7)^2 + (908 - 964,7)^2 = 745,29 + 3516,49 + 5580,09 + 28,09 + 18306,09 +$$
$$+9158,49 + 3214,89 = 40549,43$$

$$s_x^{'} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{40549,43}{6}} = 82,2$$

$t = \frac{\overline{x} - \mu_0}{s_x^{'}} \cdot \sqrt{n} = \frac{964,7 - 1000}{82,2} \cdot \sqrt{7} = -1,1362 \qquad W_a = \{t \le -t_{1-\alpha}\}$

$t_{1-0,05} = t_{0,95} = 1,943 \qquad -1,1362 > -1,943 \Rightarrow T \notin W_a \Rightarrow \textit{zamietame } H_1$

Dokázali jsme, že roznášeč letáků nerozdává méně než je dlouhodobý průměr. Počítali jsme střední hodnotu normálního rozdělení, zároveň byl malý výběr a neznali jsme rozptyl. Dali jsme si pozor na 1 stupeň volnosti u Studentova rozdělení.

3. Výrobce garantuje v jednom balíčku bonbónů obsah vitamínu C ve výši 100 mg na balení. Po změně výrobní technologie jsme vybrali 100 balíčků a naměřili průměrný obsah vitamínu 96 mg se směrodatnou odchylkou 8. Zhodnoťte, zda je výrobcovo tvrzení stále oprávněné.

$\mu_0 = 100\,mg \qquad n = 100 \qquad \overline{x} = 96\,mg \qquad s_x = 8\,mg$

$H_0 : E(X) = \mu_0 \qquad H_1 : E(X) < \mu_0 \Rightarrow W_a = \{U \le -u_{1-\alpha}\}$

$\alpha = 0,05 \Rightarrow 1 - \alpha = 0,95 \qquad u_{0,95} = 1,645$

$U = \frac{\overline{x} - \mu_0}{s_x^{'}} \cdot \sqrt{n} = \frac{96 - 100}{8} \cdot \sqrt{100} = -5$

$-5 \le -1,645 \Rightarrow \textit{platí}$

$T \in W_a \Rightarrow \textit{zamítáme } H_0;\ \textit{platí } H_1$

Po změně technologie klesl obsah vitamínu C v balení pod 100 mg s pravděpodobností 95%.

4. Před změnou vedoucího pracovníci flákali 80% pracovního času. Po změně vedoucího zjistili, že ze 400 hodin odflákali 324. Změnil se poměr odflákaného času po změně vedoucího?

Veličina je rozdělená alternativně, před změnou vedoucího se pracovníci flákali s pravděpodobností 0,8. Proto testové kritérium a kritický obor vybíráme z tabulky s parametrem π alternativního rozdělení. Za p dosazujeme novou pravděpodobnost - 324/400.

$H_0 : \pi = \pi_0$ *flákají sa tolik co předtím*

$H_1 : \pi \neq \pi_0$ *flákají se méně nebo víc*

$W_a = \{|U| \geq u_{1-\alpha/2}\}$ $u_{0,975} = 1,960$ $p = 324 / 400 = 0,81$

$$U = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{n}}} = \frac{0,81 - 0,8}{\sqrt{\frac{0,8 \cdot (1-0,8)}{400}}} = \frac{0,01}{\sqrt{0,0004}} = 0,5$$

$0,5 \geq 1,96 \Rightarrow$ *neplatí*

Zamítáme alternativní hypotézu, neprokázala se změna poměru odflákaných hodin.

5 Chí – kvadrát (χ^2) test dobré shody

Test dobré shody je takzvaný **neparametrický test**, a tady takový, že nemusíme při jeho počítání vědět, jaké rozdělení má zkoumaná veličina. χ^2 je zase jedním z neparametrických testů dobré shody. V podstatě jde o to, že o veličinách, které zkoumáme, známe nějaké předpoklady. To jsou většinou nějaké **teoretické hodnoty** – tedy statistickým nebo expertním odhadem určené velikosti veličin jako např.: *produkce ve stavebnictví bude v lednu na úrovni 75% oproti květnu, anebo: zájem o kurzy na soukromé umělecké škole se odhadují ve výši 50% klavír, 25% kytara a 25% housle*. Dále jsou známé nějaké **naměřené (empirické) hodnoty**, tedy skutečná výše stavební produkce v daných měsících, skutečné počty dětí přihlášených do školy. A **naší úlohou je určit** za použití χ^2 - testu dobré shody, **zda odchylka odhadnuté hodnoty od té skutečně naměřené vznikla jen náhodně nebo zda byl odhad špatný.**

1. *Kostkou jsme házeli 30 krát a výsledky jsou zapsané v tabulce. Rozhodněte na pravděpodobnosti 95%, zda je kostka spolehlivá.*

X	1	2	3	4	5	6
N	4	6	7	2	5	6
o	5	5	5	5	5	5

V prvním řádku je X, číslo, které padlo na kostce. Ve třetím řádku je o – předpoklad, předpokládá se totiž, že každé jednotlivé číslo padne ve třiceti pokusech právě 5krát. To by bylo ideální, přesně podle pravděpodobnosti. Ve druhém řádku je realita, tedy kolikrát které číslo padlo. Jak je vidět, ideální to není, ale naší úlohou je zjistit, jestli je to jen náhodná odchylka nebo je kostka nespolehlivá.

Když se teď podíváme na vzorec, nejdříve si všimněme hypotézy. Nulovou hypotézou bude pořád, že $\pi = \pi_0$, tzn., že předpoklad se splnil (a odchylka je náhodná), druhou hypotézou je - H_1: non H_0, a teda, že předpoklad se nesplnil (odchylka má nějakou příčinu nebo byl špatný odhad). Hypotézy jsou teda dané a můžeme vypočítat testové kritérium, které je v tomto případě χ^2:

$$\chi^2 = \sum \frac{(n-o)^2}{o}$$

přičemž **„n" je skutečnost a „o" je odhad**. Vzorec je mírně zjednodušený oproti oficiálnímu, ale výsledky jsou stejné. Jde o to, že od každé skutečné hodnoty odečteme příslušnou teoretickou hodnotu (odhad), výsledek umocníme na druhou a vydělíme teoretickou hodnotou. Na závěr všechna tato čísla sečteme a dostaneme tak definitivní výsledek – hodnotu testového kritéria χ^2, kterou porovnáváme s příslušným kvantilem χ^2 z tabulek a vyjádříme se k tomu, kterou hypotézu zamítáme, podle toho, zda platí nebo neplatí nerovnice kritického oboru. **Pozor** na stupně volnosti, řádek v tabulkách vybereme podle toho, **kolik máme x_i,** *(kolik je možností, možných odpovědí…)* avšak χ^2 má jeden stupeň volnosti, takže od toho musíme **odečíst jedničku**. V tomto případě máme 6 možností, co může padnout na kostce, takže se díváme do pátého řádku.

$H_0 : \pi_1 = \pi_2 = ... = \pi_6$ *kostka je spolehlivá*

$H_1 : non\ H_0$ *kostka je nespolehlivá*

$$T = \sum \frac{(n-o)^2}{o} = \frac{(4-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(7-5)^2}{5} + \frac{(2-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(6-5)^2}{5} = \frac{16}{5} = 3,2$$

$W_a = \{T \geq \chi^2_{1-\alpha}\}$ $\chi^2_{0,95}(5) = 11,07$

$3,2 < 11,07 \Rightarrow$ *zamítáme* H_1*, neprokázala se nespolehlivost kostky*

Jak vidíme, testové kritérium vyšlo 3,2, avšak 0,95 kvantil χ^2 rozdělení je 11,07, pro zamítnutí H_0 by muselo testové kritérium být větší než kvantil. Nejsme v kritickém oboru, zamítáme hypotézu H_1 a můžeme potvrdit hypotézu H_0 a kostku označit za spolehlivou, protože její odchylka je jen náhodná.

2. *Marketingový plán tvrdil, že záznam koncertu skupiny Kulička se prodá v poměru 50% CD, 30% DVD a 20% kazety. Za měsíc se skutečně prodá 2552 kusů CD, 923 DVD a 384 kusů kazet tohoto koncertu. Ověřte, zda byl předpoklad marketingového plánu správný.*

Nosič	CD	DVD	MC
n	2 552	923	384
o	1 930	1 158	771
n-o	622	-235	-387

$H_0 : \pi_1 = \pi_2 = \pi_3$ *předpoklad byl správný*

$H_1 : non\ H_0$ *předpoklad nebyl správný*

$$T = \sum \frac{(n_i - o_i)^2}{o_i} = \frac{622^2}{1930} + \frac{(-235)^2}{1158} + \frac{(-387)^2}{771} = 200,46 + 47,69 + 194,25 = 442,4$$

$W_a = \{T \geq \chi^2_{1-\alpha}\}$ $\chi^2_{0,95}(2) = 5,99$

$442,4 \geq 5,99 \Rightarrow$ *zamítáme* H_0*, předpoklad nebyl správný*

Nejprve si uděláme tabulku, v prvním řádku je skutečnost kolik kusů různých nosičů bylo prodaných. Když spočítáme celkový prodej a vynásobíme příslušným procentem, které bylo uváděné v marketingovém plánu, vyjde nám, jaký byl předpoklad v kusech, tj. druhý řádek tabulky. Ve třetím řádku je už jen vypočítaný rozdíl pro lepší počítání. Dosadíme čísla do vzorce a vypočítáme testové kritérium. Když budeme hledat kvantil χ^2 rozdělení v tabulkách, nesmíme zapomenout, že řádek kvantilu χ^2 odvodíme jako počet sloupců v tabulce, kterou si vytvoříme, minus 1. Pravděpodobnost není daná, tak používáme standardně 0,95,

3. *Při náhodném průzkumu bylo 25 lidí označených jako osoby malé postavy, 53 jako osoby střední postavy a 42 velké postavy. Ověřte na 5% hladině tvrzení o rovnoměrném rozdělení podle velikosti.*

„Tvrzení o rovnoměrném rozdělení podle velikosti" je jinak řečeno, že by všechny tři velikosti lidí měli být ve společnosti stejně zastoupené. Při průzkumu bylo oslovených 120 lidí, a pokud by měla mít každá ze tří velikostí stejné zastoupení, byla by třetina lidí (40) malé postavy, třetina střední a třetina velké. Z toho vychází i hodnoty odhadu, které si zapíšeme do tabulky:

postava	malá	střední	velká
n	25	53	42
o	40	40	40
n - o	-15	13	2

$$H_0 : \pi_1 = \pi_2 = \pi_3 \qquad \textit{rozdělení je rovnoměrné}$$
$$H_1 : non\, H_0 \qquad \textit{rozdělení není rovnoměrné}$$
$$T = \sum \frac{(n_i - o_i)^2}{o_i} = \frac{(-15)^2}{40} + \frac{13^2}{40} + \frac{2^2}{40} = \frac{398}{40} = 9,95$$
$$W_a = \left\{ T \geq \chi^2_{1-\alpha} \right\} \qquad \chi^2_{0,95}(2) = 5,99$$
$$9,95 \geq 5,99 \Rightarrow \textit{zamítáme } H_0, \textit{rozdělení není rovnoměrné}$$

4. Na základě údajů z tabulky určete, zda kvalita výrobků závisí na tom, která směna je vyráběla.

počty výrobků	jakost I.	jakost II	jakost III
vyrobených za dopolední směnu	170	250	80
vyrobených za odpolední směnu	160	300	60

Kvalita vyráběných výrobků je daná počtem výrobků první, druhé a třetí jakosti, které se během směny vyrobí. Pochopitelně, v různých směnách mohou pracovat různé počty lidí, kteří vyrobí celkově různé množství výrobků, proto se tyto údaj nedají přímo porovnat. Co bude teda odhadem, odhadovaným množstvím výroby? To bude takové množství, které by dokazovalo, že kvalita nezávisí na směně. Pokud kvalita nezávisí na směně, která výrobky vyrábí, měly by obě směny vyrábět stejné množství každé jakosti. Tady se ale dostáváme k problému, že jedna směna vyrobí víc výrobků než druhá. Proto musí být odhadované množství vážené celkovým množstvím výroby dané směny. Výpočet odhadovaného množství výrobků první jakosti uděláme tak, že celkový počet výrobků první jakosti (330) vydělíme celkovým počtem vyrobených výrobků (1020), tím pádem dostaneme podíl I. jakosti na vyrobených součástkách. Tento podíl potom vynásobíme počtem všech součástek vyrobených první směnou (500) a dostaneme odhadovaný počet výrobků první jakosti pro první směnu. Podobně vynásobíme tento podíl počtem všech součástek druhé směny (520). A potom pokračujeme s druhou jakostí atd. To, že dělíme množství výroby určité jakosti 1020 a potom násobíme 500 a 520 nám zabezpečí rovnoměrné rozdělení dané jakosti s ohledem na množství výroby během směny (nás nezajímá množstvní, ale jen kvalita).

$$odhad\ (I.akost,\ dopoledne) = \frac{(170+160)}{1020} \cdot 500 = 161,8$$
$$odhad\ (I.akost,\ odpoledne) = \frac{(170+160)}{1020} \cdot 520 = 168,2$$
$$\sum (I.\ akost) = 161,8 + 168,2 = 330$$

Podobně dopočítáme odhady pro všechny směny a jakosti a už nám zůstává jen vypočítané odhady porovnat s realitou.

Směna/kvalita	jakost I	jakost II	jakost III	$\sum$
dopoledne	170	250	80	500
odpoledne	160	300	60	520
$\sum$	330	550	140	1020
odhad dopoledne	161,8	269,6	68,6	500
odhad odpoledne	168,2	280,4	71,4	520

$H_0 : \pi_1 = \pi_2 = ... = \pi_6$ *kvalita nezávisí na směně*

$H_1 : non\, H_0$ *kvalita závisí na směně*

$$T = \sum \frac{(n_i - o_i)^2}{o_i} = \frac{(170-161,8)^2}{161,8} + \frac{(160-168,8)^2}{168,8} + \frac{(250-269,6)^2}{269,6} + \frac{(300-280,4)^2}{280,4} +$$

$$+ \frac{(80-68,6)^2}{68,6} + \frac{(60-71,4)^2}{71,4} = 0,416 + 1,425 + 1,894 + 0,4 + 1,37 + 1,82 = 7,425$$

$W_a = \{T \geq \chi^2_{1-\alpha}\}$ $\chi^2_{0,95}(2) = 5,99$

$7,425 \geq 5,99 \Rightarrow$ *zamítáme* H_0, *kvalita závisí na směně*

5. Otestuj na 5% hladině významnosti předpoklad o nezávislosti odpovědí na pohlaví:

Pohlaví/Odpověď	Ano	Ne
Muž	25	40
Žena	35	40

Odhady počtu odpovědí v případě, že by nebyly závislé na pohlaví, musíme stanovit podobně jako v předešlém příkladě. Vidíme, že žen bylo v anketě více než mužů, počet odpovědí musí být vyrovnaný, avšak vážený v prospěch žen. Takže odpovědi ANO dohromady (25+35=60) dělené celkovým počtem odpovědí (140) a nakonec vynásobit celkovým počtem mužů a celkovým počtem žen.

pohlaví/odpověď	ano	ne	$\sum$
muž	25	40	65
žena	35	40	75
Σ	60	80	130
muž	27,86	37,14	65
žena	32,14	42,86	75

6 Analýza rozptylu (ANOVA)

Analýzu rozptylu využíváme v případě, že máme **víc druhů testovaného předmětu, a zároveň každý druh testujeme vícekrát**. Například máme tři druhy benzínu a testujeme jeho spotřebu v pěti jízdách pro každá druh. Je jasné, že každá jízda se bude ve spotřebě lišit, protože spotřebu paliva ovlivňuje mnoho faktorů. Každopádně se bude spotřeba v pěti jízdách pro jeden druh benzínu točit okolo nějaké střední hodnoty a bude mít nějaký rozptyl. Tento rozptyl spotřeby benzínu jednoho druhu v pěti jízdách, rozptyl jedné skupiny, se nazývá **vnitroskupinový rozptyl** - $S_{y,v}$. Jednotlivé skupiny (druhy benzínů) mají svůj celkový průměr, který jsme vypočítali z těch pěti měření. Ale i tyto průměry se budou u každé skupiny pravděpodobně lišit a můžeme stanovit jejich střední hodnotu a rozptyl. Tento rozptyl, protože nám něco říká o kolísání hodnot mezi jednotlivými druhy benzínu, mezi jednotlivými skupinami se nazývá **meziskupinový rozptyl** - $S_{y,m}$. A nakonec **celkový rozptyl** – tedy mezi skupinami i uvnitř skupin se vypočítá jako jednoduchý součet: $S_y = S_{y,v} + S_{y,m}$. Pokud počítáme příklady, jedná se většinou o **testování hypotézy, že střední hodnoty skupin se rovnají**. Takže nulová hypotéza bude předpokládat, že nejsou rozdíly mezi jednotlivými druhy, proti čemuž postavíme alternativní hypotézu, která popírá nulovou – tedy, že tam nějaký rozdíl je, že alespoň jedna střední hodnota se odlišuje a není to jen statistická odchylka. Testovým kritériem je F-test, jehož výsledek zase jen porovnáme s kvantilem z tabulek a podle toho, zda proměnná patří do kritického oboru, zamítneme jednu z hypotéz.

1. *Zkoumali jsme tři druhy benzínu a u každého jsme udělali 5 měření spotřeby. Meziskupinový rozptyl jsme vyčíslili na 0,25 a vnitroskupinový se rovná 0,08. Ověřte hypotézu, že se spotřeby u těchto třech druhů rovnají.*

$S_{y,m} = 0,25 \qquad S_{y,v} = 0,08 \qquad k = 3 \qquad n = 15$

$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad W_a = \{F \geq F_{1-\alpha}\}$

$H_1 : non\ H_0 \qquad F_{0,95}(2;12) = 3,885$

$$F = \frac{\frac{S_{y,m}}{k-1}}{\frac{S_{y,v}}{n-k}} = \frac{\frac{0,25}{2}}{\frac{0,08}{12}} = 18,75 \qquad 18,75 \geq 3,885 \Rightarrow \textit{platí}$$

Zamítáme hypotézu H_0.

Proměnnou *k* je počet skupin – v našem případě 3 – a proměnnou *n* je počet měření. Pozor, proměnná „počet měření" je myšlená jako **součet počtu měření v každé skupině**. V našem případě v každé ze skupin bylo 5 měření, tedy celkový počet měření (n) je 5+5+5=15. Jediný problém snad může nastat při hlednání hodnoty kvantilu v tabulkách, protože F-rozdělení má stupně volnosti. Sloupec v tabulce určíme vypočítáním $\upsilon_1 = k - 1$. Řádek, ve kterém najdeme hledanou hodnotu, určíme vypočítáním $\upsilon_2 = n - k$.

2. *Zkoumali jsme odrůdy brambor a udělali dohromady 28 měření. Z nich jsme zjistili, že vnitroskupinový rozptyl je dvakrát větší než meziskupinový a testové kritérium nám vyšlo F=4. Zjistěte, kolik odrůd jsme zkoumali na hladině pravděpodobnosti $\alpha = 0,05$ zamítneme nulovou hypotézu.*

$$F = 4 \qquad n = 28 \qquad S_{y,v} = 2 \cdot S_{y,m} \quad k = ?$$

$$F = \frac{\frac{S_{y,m}}{k-1}}{\frac{S_{y,v}}{n-k}} \qquad 4 = \frac{\frac{S_{y,m}}{k-1}}{\frac{2 \cdot S_{y,m}}{28-k}}$$

$$4 = \frac{S_{y,m}}{k-1} \cdot \frac{28-k}{2 \cdot S_{y,m}}$$

$$4 = \frac{1}{k-1} \cdot \frac{28-k}{2} = \frac{28-k}{2k-2}$$

$$8k - 8 = 28 - k$$

$$9k = 36$$

$$k = 4$$

Jednoduchými matematickými úpravami a substitucí vyplývající ze vztahu meziskupinového a vnitroskupinového rozptylu jsme přišli k počtu zkoumaných odrůd. Nyní můžeme potvrdit nebo vyvrátit hypotézu. Testové kritérium máme zadané, stačí jen najít správný kvantil F-rozdělení.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \qquad W_a = \{F \le F_{1-\alpha}\}$$

$$H_1 : non\ H_0$$

$$F_{0,95}(k-1; n-k) = F_{0,95}(3; 24) = 3,008$$

$$4 \le 3,008 \Rightarrow neplatí$$

Hodnota testového kritéria nepatří do kritického oboru – zamítám hypotézu H_1. Rozdíl ve středních hodnotách naměřených veličin se neprokázal.

3. *Dopočítejte chybějící hodnoty do tabulky, udělejte test včetně zapsání hypotéz a vyvoďte závěry, vhodným ukazatelem změřte sílu závislosti a závěr též okomentujte.*

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	456, 1333333	228, 0666	?	<.0001
Error	12	23, 33333333	?		
Corrected Total	14	?			

Co se týká doplňování hodnot do tabulky, pomůže nám vysvětlující tabulka:

Zdroj	Stupně volnosti	Součet čtverců	Průměr čtverců	Testové kritérium F	P - hodnota
Model	k-1	Sy,m	Sy,m/(k-1)	F	P - hodnota
Error	n-k	Sy,v	Sy,v/(n-k)		
Corrected Total	n-1	Sy			

Když víme, které číslo co znamená, neměl by být problém doplňovat čísla. První otazník (zleva) doplníme na základě vztahu $S_y = S_{y,v} + S_{y,m}$ = 479,466. Druhý otazník dostaneme vydělením součtu čtverců stupni volnosti = 23,3333/12 = 1,9444. Třetí otazník je testové kritérium, které musíme vypočítat:

$$F = \frac{\frac{S_{y,m}}{k-1}}{\frac{S_{y,v}}{n-k}} = \frac{\frac{456,1333}{2}}{\frac{23,3333}{12}} = \frac{228,0666}{1,9444} = 117,294$$

Dále máme udělat test, neboli stanovit hypotézy a porovnat vypočítanou hodnotu testového kritéria F s kritickým oborem:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad F = 117,294$$
$$H_1 : non\ H_0 \qquad W_a = \{F \geq F_{1-\alpha}\}$$
$$F_{1-\alpha}[k-1; n-k] = F_{0,95}(2;12) = 3,885$$
$$117,294 \geq 3,885 \Rightarrow \textit{platí}$$

Testové kritérium patří do kritického oboru, proto zamítáme nulovou hypotézu a konstatujeme, že střední hodnoty zkoumaných předmětů se liší. Změřit sílu závislosti můžeme koeficientem determinace (podrobněji viz strana 44):

$$P^2 = \frac{S_{y,m}}{S_y} = \frac{456,1333}{479,4666} = 0,95133$$

Dokázali jsme silnou závislost a model můžeme označit za kvalitní.

7 Regresní a korelační analýza

Regrese je **závislost jedné veličiny na druhé**. Například množství žen v sukni na Václavském náměstí v závislosti na teplotě vzduchu. Závislost je tam jasná, čím vyšší teplota, tím víc žen v sukni. Teplota vzduchu se v tomto případě nazývá **vysvětlující** proměnná, protože její změna nám vysvětluje změnu jiné proměnné. A tou je počet žen v sukni – ten je **vysvětlovanou** proměnnou, tzn. proměnnou, jejíž změnu chápeme jako důsledek změny vysvětlující proměnné.

7.1 Regresní přímka

Závislost dvou proměnných v regresní analýze se snažíme **popsat nějakou funkcí či křivkou**. To znamená, že najdeme takovou funkci, která pro každou hodnotu vysvětlující proměnné (standardně označujeme jako X) najde co nejpravděpodobnější hodnotu vysvětlované proměnné (označujeme Y). Když takovou funkci graficky vyjádříme, bude to nějaká křivka. Vzhledem k tomu, jakým způsobem jsou na sobě proměnné závislé, mění se i vzhled křivky a druh funkce. Nejzákladnějším a zároveň i nejběžnějším způsobem, jakým mohou být na sobě dvě proměnné závislé, je **lineární závislost**. Ta může být buď **přímá** (čím víc stlačím plynový pedál, tím rychleji jedu), anebo **nepřímá** (čím je auto starší, tím je menší jeho hodnota). Pro lineární závislost je charakteristické, že je jedno, jak velké je x, změna o jednu jednotku způsobí vždy stejnou změnu y (např. když zvýším proměnnou x z 10 na 11, zvýší se proměnná y o 10, když zvýším x z 243 na 244, y se zvýší o 10). Grafickým vyjádřením lineární závislosti je regresní přímka. Matematickým vyjádřením je funkce

$$y = \beta_0 + \beta_1 x$$

která má dva parametry β_0 a β_1, přičemž vidíme, že β_0 je jakási pevná složka *(auto bude něco stát, ať bude jakkoli staré) a* β_1 se bude měnit v závislosti na x. Zároveň je jasné, že pokud bude β_1 kladná, tak se zvyšujícím se x se bude zvyšovat i y, půjde tedy o přímou závislost, naopak bude-li β_1 záporná, půjde o závislost nepřímou. Otázka je, jak se dopracovat k hodnotám těchto parametrů. Na jejich určení použijeme metodu nejmenších čtverců, přičemž jako první parametr vypočítáme β_1 a β_0 potom jednoduše dopočítáme použitím hodnoty β_1. Nejlépe to ukáže příklad:

1. *Zaměstnanci firmy se zapracovávají na nové výrobní lince. Pro šest zaměstnanců je zaznamenávaný počet doteď odpracovaných hodin (veličina X) a zjištěný procentuální podíl chybných výrobků (veličina Y). Určete regresní přímku (tj. hodnoty jejích parametrů) závislosti Y na X a interpretujte co nejvýstižněji hodnotu směrnice.*

Zaměstnanec č.	1	2	3	4	5	6
Odprac. hod.	82	86	87	87	91	95
% chyb	11	10	12	9	10	8

Vidíme, že počet chybných výrobků by měl záviset na odpracovaných hodinách na nové lince. Označíme tedy počet odpracovaných hodin jako X (vysvětlující, nezávislá proměnná) a počet chyb jako Y (vysvětlovaná, závislá proměnná, její velikost závisí na velikosti X). Když máme jasno v proměnných, můžeme přistoupit k počítání koeficientů funkce.

Koeficient β_1 vypočítám pomocí metody nejmenších čtverců, používám vzorec:

$$\beta_1 = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - \overline{x}^2}$$

Jak vidíme, budeme si muset nejprve připravit proměnné do tohoto vzorce, vypočítáme průměry X,Y, součin jejich průměrů, průměr druhých mocnin X a druhou mocninu průměru X.

Zaměstnanec č.	1	2	3	4	5	6	**Průměr**
Odprac. hod. (X)	82	86	87	87	91	95	88
% chyb (Y)	11	10	12	9	10	8	10
X^2	6724	7396	7569	7569	8281	9025	7760,67
X.Y	902	860	1044	783	910	760	876,5

Teď už máme všechny proměnné a můžeme je dosadit do vzorce

$$\beta_1 = \beta_{xy} = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{876,5 - (88 \cdot 10)}{7760,67 - 88^2} = \frac{-3,5}{16,67} = -0,21$$

Máme vypočítaný koeficient β_1 (který se nazývá i **regresní koeficient** a často je označovaný jako β_{xy}) a můžeme dopočítat β_0 podle vzorce:

$$\beta_0 = \overline{y} - \beta_1 \cdot \overline{x}$$

$$\beta_0 = \overline{y} - \beta_1 \cdot \overline{x} = 10 - (-0,21) \cdot 88 = 10 + 18,48 = 28,48$$

Směrnicí regresní přímky je hodnota koeficientu β_1, přičemž na základě toho, zda je její znaménko plus nebo mínus, umíme určit, jestli je závislost přímá nebo nepřímá. V tomto případě je před směrnicí přímky **znaménko mínus, což značí nepřímou závislost**. Takže čím víc hodin pracovníci na lince odpracují, tím méně chyb dělají. Jako výsledek zapíšeme hotovou rovnici regresní přímky ve tvaru $y = \beta_0 + \beta_1 x$, a vyjádříme se o typu závislosti:

Závislost chybovosti na počtu odpracovaných hodin na stroji je lineární, daná regresní přímkou $y = 28,48 - 0,21 \cdot x$, přičemž hodnota směrnice nám udává, že jde o nepřímou závislost.

7.2 Těsnost závislosti

V minulém příkladě jsme vyjádřili závislost vysvětlované proměnné Y na vysvětlující proměnné X. Tato závislost však může mít různou sílu, to znamená, že naše přímka (či funkce), kterou jsme tuto závislost popsali, buď přesně odpovídá realitě nebo má od ní nějaké odchylky nebo vůbec nesedí a celá ta funkce, ke které jsme se dopracovali, by byla na nic. Sílu závislosti můžeme měřit vícero koeficienty. **Index determinace** dává například do poměru **teoretický rozptyl a celkový rozptyl.** Čím větší je podíl teoretického rozptylu na rozptylu celkovém, tím je závislost silnější a použitý model tuto závislost lépe vystihuje (podrobněji viz strana 55).

$$I_{yx}^2 = \frac{s_T^2}{s_y^2} = \frac{\sum (Y_i - \overline{y})^2}{\sum (y_i - \overline{y})^2} \qquad s_T^2 = teoretický\ rozptyl \qquad s_y^2 = celkový\ rozptyl$$

Index determinace je založený na tom, že čím je větší podíl teoretického rozptylu na celkovém rozptylu, tím je závislost silnější, těsnější a zároveň nám říká o tom, nakolik je námi zvolená funkce správná pro popis reality. **Jeho hodnota se pohybuje v intervalu <0;1>**, přičemž 1 znamená dokonalou závislost a zároveň použitou funkci za absolutně vhodnou, v případě 0 znamená buď žádnou závislost, nebo nesprávně zvolenou funkci.

Samotný **index determinace se však většinou v praxi nepoužívá k určování těsnosti závislosti**, ale používá se odmocnina z něj – **index korelace**:

$$I_{yx} = \sqrt{\frac{s_T^2}{s_y^2}} = \sqrt{\frac{\sum (Y_i - \overline{y})^2}{\sum (y_i - \overline{y})^2}}$$

Z tohoto indexu vyvozujeme závěry stejně jako z indexu determinace.

V případě, že se jedná o lineární regresi (křivkou, která popisuje závislost je přímka a funkce popisující závislost je ve tvaru $\beta_0 = \overline{y} - \beta_1 \cdot \overline{x}$), můžeme matematickými úpravami dospět k **jednoduššímu vzorci indexu korelace** pro výpočet, u kterého už nepotřebujeme všechny hodnoty, ale stačí nám jen souhrnné statistiky jako průměry a druhé mocniny průměrů. Nazývá se **korelační koeficient**:

$$r_{yx} = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sqrt{\left(\overline{x^2} - \overline{x}^2\right)\left(\overline{y^2} - \overline{y}^2\right)}}$$

Tento koeficient **nabývá hodnoty od -1 do 1** a podobně jako směrnice funkce (β_1) nám znamínkem říká, zda je závislost přímá nebo nepřímá. Zároveň ale, **čím je blíž nule, tím je závislost slabší**, v případě nuly není mezi X a Y žádná lineární závislost.

2. Určete druh a sílu závislosti proměnné Y na X, když jsme naměřili tyto hodnoty.

x	12	5	6	10	14	8	9	12
y	8	4	4	8	11	6	5	10

Stačí nám spočítat korelační koeficient, ze kterého vyčteme druh i sílu závislosti.

									průměr
X	12	5	6	10	14	8	9	12	$\overline{x}$ =9,5
Y	8	4	4	8	11	6	5	10	$\overline{y}$ =7
X.Y	96	20	24	80	154	48	45	120	$\overline{xy}$ =73,375
X^2	144	25	36	100	196	64	81	144	$\overline{x^2}$ =98,75
Y^2	64	16	16	64	121	36	25	100	$\overline{y^2}$ =55,25

$$r_{yx} = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sqrt{\left(\overline{x^2} - \overline{x}^2\right).\left(\overline{y^2} - \overline{y}^2\right)}} = \frac{73,375 - (9,5 \cdot 7)}{\sqrt{(98,75 \cdot 9,5^2)(55,25 \cdot 7^2)}} = \frac{6,875}{7,289} = \underline{0,943}$$

Hodnota korelačního koeficientu ukazuje na velmi silnou přímou závislost.

Navzdory tomu, že nám korelační koeficient ukáže silnou závislost, nemusí to nic znamenat při malém počtu hodnot, ze kterých jsme vycházeli (málo x a y). Proto si **můžeme testem ještě ověřit, jestli jsou proměnné skutečně závislé** tak, jak ukazuje korelační koeficient. Testujeme jako nulovou hypotézu, že se korelační koeficient rovná nule – to by značilo nezávislost proměnných, jako alternativní hypotézu volíme non H0.

3. *Měřili jsme 10 hodnot a stanovili lineární regresní funkci, přičemž její absolutní člen se rovná 2354, regresní koeficient je roven -5438 a index determinace je 56%. Určete regresní funkci, typ a sílu závislosti a udělejte test o korelačním koeficientu na hladině pravděpodobnosti* $1-\alpha = 0,95$*.*

Nejprve sestavíme regresní funkci. Absolutní člen je jenom jiné pojmenování pro β_0 a regresní koeficient je zase β_1. Výsledná funkce bude tedy ve tvaru:

$$y = 2354 - 5438x$$

Teď se potřebujeme dostat ke korelačnímu koeficientu. Musíme si uvědomit, že korelační koeficient je jen jednodušším vyjádřením indexu korelace. Máme zadaný **index determinace a z toho dostaneme index korelace odmocněním**. Budeme proto pracovat s hodnotou korelačního koeficientu (indexu korelace) $\sqrt{0,56} = 0,748$, která nám říká, že závislost Y na X je středně silná a přímá. My ale uděláme ještě test o korelačním koeficientu:

$$H_0 : \rho_{yx} = 0 \qquad t = \frac{r_{yx}\sqrt{n-2}}{\sqrt{1-r_{yx}^2}} = \frac{0,748 \cdot \sqrt{8}}{\sqrt{1-0,748^2}} = \frac{2,11566}{0,6637} = 3,18767$$

$$H_1 : \rho_{yx} \neq 0$$

$$W_a = \{|t| \geq t_{1-\alpha/2}\} \qquad t_{0,95} = 1,86 \qquad 3,18767 > 1,86$$

Vypočítané testové kritérium je větší než kvantil Studentova rozdělení, proměnná patří do kritického oboru. Zamítáme hypotézu H0 a konstatujeme, že mezi X a Y existuje závislost. Při Studentově rozdělení pozor na stupně volnosti – musíme se dívat do n-2 řádku v tabulkách.

Kromě lineární závislosti můžeme samozřejmě narazit i na závislosti jiných druhů. Složitější závislosti jsou popsané složitějšími funkcemi, například parabolickou, hyperbolickou nebo polynomickou. V případě paraboly, která je jedinou z těchto složitějších regresí, kterou se budeme zabývat, má regresní funkce tvar: $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Častou úlohou je například rozhodnout, která funkce lépe opisuje danou skutečnost, přičemž v praxi jsou stále preferované co nejjednodušší modely, protože se s nimi snadněji počítá a snáz se interpretují výsledky. **Při rozhodování o tom, která funkce je vhodnější**, **budeme používat index determinace**:

$$I_{yx}^2 = R^2 = \frac{S_T}{S_y}$$

Na to, abychom jej vypočítali, je nutné znát buď všechny hodnoty, ze kterých jsme dané funkce odvodili nebo nám stačí znát teoretický a celkový rozptyl.

4. U deseti zkoumaných jednotek byly zjištěny hodnoty proměnných a z nich vypočítány tyto trendové funkce:
Y = 55 + 12,5 x
Y = 130 + 60,071 x + 0,107 x²

a tyto součty odchylek čtverců:

Analysis of variance				
	Pro přímku		Pro parabolu	
Source	DF	Sum of Squares	DF	Sum of Squares
Model	1	312 500	2	315 714
Error	8	103 500	7	100 286
Corrected Total	9	416 000	9	416 000

Vyberte vhodnější regresní funkci a spočítejte hodnotu proměnné Y pro x=200.

Hodnoty rozptylů (odchylek čtverců), které potřebujeme, máme uvedené v tabulce. V řádku „Model" máme teoretický rozptyl S_T a v řádku „Corrected total" je rozptyl celkový - S_y. Hodnota v řádku „Error" se nazývá reziduální rozptyl S_R a pro tyto tři rozptyly platí vztah:

$$S_y = S_R + S_T$$

Vypočítáme index determinace za použití vzorce a hodnot z tabulky:

$$pro\ \ přímku: \qquad R^2 = \frac{S_T}{S_y} = \frac{312\,500}{416\,000} = 0,7512$$

$$pro\ \ parabolu: \qquad R^2 = \frac{S_T}{S_y} = \frac{315\,714}{416\,000} = 0,7589$$

Pokud bychom nepokračovali dál a rozhodli se podle velikosti indexu determinace (například bychom neměli k dispozici údaj o stupních volnosti), tak bychom si podle výpočtů měli vybrat jako vhodný model parabolu, protože má vyšší index determinace. To ale není úplně správné, obecně jsou preferované modely, které jsou jednodušší, takže lineární model je preferovanější než parabolický. V případě takto malého rozdílu bychom se tedy rozhodli pro lineární model. My ale budeme pokračovat a vypočítáme si **modifikovaný index determinace**. Ten by vytvořený z důvodu, že obyčejný index determinace favorizuje modely, které mají více proměnných. Čím víc proměnných v modelu, tím je vyšší index determinace. Tuto deformaci odstraníme tak, že snížíme velikost indexu determinace o takovou hodnotu, o jakou byl zvýšený v důsledku vyššího počtu proměnných. Modifikovaný index determinace dostaneme úpravou klasického:

$$I^2_{ADJ} = 1-(1-I^2_{yx})\cdot\frac{n-1}{n-p}$$

Ve vzorci nacházíme hodnotu (n-1) – což je hodnota ve sloupci DF a v řádku Corrected Total, takže stupně volnosti celkového rozptylu. Za (n-p) zase dosadíme hodnotu sloupce DF a v řádku Error – tedy stupně volnosti reziduálního součtu čtverců. V proměnné p se v podstatě skrývá počet proměnných použitého modelu.

$$pro\ přímku: \quad R^2_{adj} = 1-(1-I^2_{yx})\cdot\frac{n-1}{n-p} = 1-(1-0,7512)\cdot\frac{9}{8} = 1-0,2799 = 0,7201$$

$$pro\ parabolu: \quad R^2_{adj} = 1-(1-I^2_{yx})\cdot\frac{n-1}{n-p} = 1-(1-0,7589)\cdot\frac{9}{7} = 1-0,31 = 0,69$$

Na základě modifikovaného indexu determinace už jasně vidíme, že přímka je pro tento případ lepším modelem. Modifikovaný index determinace **je vždy o něco nižší** než nemodifikovaný.

5. *Deset letních dnů byl zaznamenávaný počet hodin, během kterých svítilo slunce a počet prodaných litrů zmrzliny (=Y). Vypočítejte hodnotu korelačního koeficientu závislosti y na x a interpretujte jeho význam. Z dat byly už vypočítány údaje v následující tabulce:*

veličina	x	y	x^2	y^2	xy
součet hodnot	25	140	72	2550	460

$$\overline{xy} = \frac{460}{10} = 46 \qquad \overline{x} = \frac{25}{10} = 2,5 \qquad \overline{y} = \frac{140}{10} = 14 \qquad \overline{x^2} = \frac{72}{10} = 7,2 \qquad \overline{y^2} = \frac{2550}{10} = 255$$

$$r_{yx} = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sqrt{\left(\overline{x^2} - \overline{x}^2\right).\left(\overline{y^2} - \overline{y}^2\right)}} = \frac{46 - 2,5\cdot 14}{\sqrt{(7,2-2,5^2)\cdot(255-14^2)}} = \frac{11}{\sqrt{0,95\cdot 59}} = 1,47$$

Korelační koeficient nemůže nabýt hodnoty vyšší než 1, proto musíme konstatovat, že zadání je špatně.

6. *Pro X a Y jsme zaznamenali hodnoty v tabulce. Pomocí regresní přímky zjistěte střední hodnotu proměnné Y, v případě kdy X=22.*

x	3	5	6	8	11	10
y	4	8	7	12	18	18

Nejprve si musíme dopočítat potřebné hodnoty druhých mocnin a průměry:

	Hodnoty						průměr
X	3	5	6	8	11	10	43/6=7,17
Y	4	8	7	12	18	18	67/6=11,17
X^2	9	25	36	64	121	100	355/6=59,17
XY	12	40	42	96	198	180	568/6=94,67

A potom dosadit správně do vzorců a vypočítat parametry regresní přímky a nakonec do ní dosadit hodnotu x=22 a vypočítat rovnici:

$$\beta_1 = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{94,67 - 7,17 \cdot 11,17}{59,17 - 7,17^2} = \frac{14,6}{7,76} = 1,88$$

$$\beta_0 = \overline{y} - \beta_1 \cdot \overline{x} = 11,17 - 1,88 \cdot 7,17 = -2,31$$

$$Y = \beta_0 + \beta_1 \cdot x \Rightarrow \underline{Y = 1,88x - 2,31}$$

$$x = 22 \Rightarrow Y(22) = 1,88 \cdot 22 - 2,31 = 39,05$$

Musíme si uvědomit, že tím, že vypočítáme rovnici konkrétní regresní přímky, jsme vlastně dostali nástroj, který nám, na základě vztahu, který mezi proměnnými je, přiřadí každému X příslušné Y nebo naopak. Naší úlohou v tomto případě bylo tedy vypočítat regresní přímku, potom do ní dosadit za X 22, a tak zjistit hodnotu Y.

8 Časové řady

Časová řada je **souhrn více měření nějaké veličiny v různém čase**, například vývoj HDP za období 2004-08, počet zkonzumovaných piv denně během jednoho týdne a podobně. Dělí se podle toho, zda je zaznamenáváme v bodě (teplota ve 12:00) nebo v intervalu (spotřeba za rok), podle toho, zda jsou krátkodobé nebo dlouhodobé a ještě podle několika kritérií. Užitečnost časové řady spočívá v tom, že ukazuje určitý trend, tedy směr a rychlost, kterými se veličina vyvíjí. A když najdeme funkci, která popisuje pohyb velikosti veličiny v minulosti, budeme ji umět využít i na předpovídání budoucnosti. A to je na tom to nejlepší. Je tu ale i dost problémů, trend, který sledujeme je stále ovlivněný mnoha faktory jako například **sezónní vlivy** (v zimě je zpravidla menší zaměstnanost jako v létě), **cyklické vlivy** (střídání recese a expanze v ekonomice), ale i **náhodné vlivy** způsobené množstvím malých, nezávislých veličin, které působí na sledovaný jev.
Při praktickém počítání s časovými řadami, i když jen v rámci jednoho roku už narážíme na problém. A to z toho důvodu, že **měsíce nemají stejný počet dní**, a i když mají, tak nemusí být stejný počet pracovních dní. Proto jsou údaje z jednoho měsíce těžko porovnatelné s údaji z měsíce jiného. Časovou řadu proto musíme očistit od takových vlivů a přepočítat veličinu na nějaký průměrný počet dní. Když očišťujeme intervalovou časovou řadu, přepočet hodnoty veličiny nepravidelného období na průměrný počet dní nám zabezpečí vzorec:

$$y_t^o = y_t \frac{\overline{k}_t}{k_t},$$

Takže – očištěnou hodnotu veličiny dostaneme tak, že její původní velikost vynásobíme podílem průměrného počtu dní (např. průměrný počet dní měsíce v roce= 365/12 = 30,42 dne) a počtu dní v daném období (kdyby byly údaje např. za březen, tak by to bylo 31). Po očištění už je možné tyto údaje porovnávat a případné odchylky zdůvodnit, když už je nebude mít na svědomí kalendář.

V případě veličin měřených bodově je užitečnou informací o vývoji **průměr**. Například když měříme teplotu každý červencový den ve dvanáct, určitě nás zajímá průměrná teplota v červenci. Pokud máme čísla naměřená ve stejných intervalech, můžeme použít **prostý chronologický průměr**:

$$\overline{y} = \frac{\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + \frac{y_3 + y_4}{2} + \ldots + \frac{y_{k-1} + y_k}{2}}{k - 1}$$

Ypsilon jsou samozřejmě hodnoty jednotlivých období a k je celkový počet období, které jsme měřili. V případě, že nemáme stejné časové rozestupy mezi naměřenými hodnotami, nemůžeme použít tento vzorec, ale musíme používat vzorec **váženého chronologického průměru** ve tvaru:

$$\overline{y} = \frac{\frac{y_1 + y_2}{2} \cdot d_1 + \frac{y_2 + y_3}{2} \cdot d_2 + \frac{y_3 + y_4}{2} \cdot d_3 + \ldots + \frac{y_{k-1} + y_k}{2} \cdot d_{k-1}}{d_1 + d_2 + d_3 + \ldots + d_{k-1}}$$

Přičemž proměnná d v tomto vzorci značí délku intervalu, časový rozestup mezi dvěma měřenými hodnotami.

1. *Podnik má k daným datům počet zaměstnanců uvedený v tabulce.*

Datum	Počet zaměstnanců
31.1.2008	645
30.4.2008	580
30.9.2008	621
31.12.2008	500

Jaký je průměrný stav zaměstnanců za období od 31.1 do 31.12. jedná-li se o přestupný rok?

Budeme si muset spočítat rozestupy mezi jednotlivými hodnotami ve dnech, jelikož nejsou stejné. Od 31.12. do 30.4. to je 29+31+30 = 90 dní, od 30.4. do 30.9. to je 31+30+31+31+30= 153 dní a od 30.9. do 31.12. to je 31+30+31 = 92 dní. A teď to už jen dosadíme do vzorce.

$$\bar{y} = \frac{\frac{645+580}{2}\cdot 90 + \frac{580+621}{2}\cdot 153 + \frac{621+500}{2}\cdot 92}{90+153+92} = \frac{55125+91876,5+51566}{335} = 592,74$$

Průměry nám naznačují stav v určitém období a jsou-li správně očištěné, dovolují nám období porovnávat. Když však chceme vědět, jaký byl v rámci období růst, tedy jak rychle se měřená veličina zvyšovala nebo snižovala, musíme použít jiné nástroje. Absolutním vyjádřením změny během sledovaného období je **absolutní přírůstek** Δ_t, jehož výpočet spočívá jen v tom, že odečteme od poslední hodnoty časové řady první hodnotu. Relativním vyjádřením je zase **relativní přírůstek** δ_t, který se vypočítá tak, že od poslední hodnoty řady odečteme první hodnotu řady a výsledek ještě vydělíme první hodnotou řady. Relativní přírůstek nám tak v podstatě udává procentuální změnu (přičemž jeho hodnota 1 = 100%) oproti první hodnotě řady. Procentuální změna je však vyjádřená souhrnně za celé období, což znamená, že když analyzujeme data od roku 2004 do roku 2008, ve výsledku bychom se dopočítali k tomu, o kolik procent je hodnota v roce 2008 větší/menší oproti hodnotám v roce 2004. Pokud ale chceme vědět, jaký byl průměrný růst každý rok (každou jednotku času) ve zkoumaném období, použijeme koeficient $\bar{k}$, to je **průměrné tempo růstu** časové řady.

$$\bar{k} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

Jak vyplývá ze vzorce, je to n-1 odmocnina z podílu posledního a prvního členu řady, kterou zkoumáme. To „n-1" znamená, že spočítáme, kolik členů má časová řada, odečteme od toho 1 a výsledkem to budeme odmocňovat.

2. *Následující tabulka uvádí výši HDP ČR na obyvatele v běžných cenách. Spočítejte průměrný koeficient růstu za období 95-98, 98-01, 01-07 a na základě koeficientu z posledního období odhadněte výšku HDP na obyvatele v letech 2008 a 2009.*

Rok	HDP
1995	141 957
1998	193 929
2001	230 064
2007	341 989

Pokud chceme spočítat průměrný koeficient růstu, první, co si musíme uvědomit, je, kolik roků má vlastně období, pro které ho chceme spočítat. HDP se udává ke konci roku a je to tedy údaj za uplynulý rok. Takže období 95-98 bude v sobě obsahovat 4 roky – 95, 96, 97, 98. Za období 98-01 to budou zase 4 roky – 98, 99, 00, 01. Zní to triviálně, ale je to asi jediné zrádné místo v příkladech tohoto typu. Zbytek je už jen dosazování do vzorce.

$$\bar{k}_{95-98} = \sqrt[3]{\frac{193929}{141957}} = 1{,}109 \qquad \bar{k}_{98-01} = \sqrt[3]{\frac{230064}{193929}} = 1{,}059$$

$$\bar{k}_{01-07} = \sqrt[6]{\frac{341989}{230064}} = 1{,}068$$

Na dopočítání hodnoty let 2008 a 2009 nám stačí považovat $\bar{k}$ za dané a y_n za neznámé.

$$1{,}068 = \sqrt[1]{\frac{y_{2008}}{341989}} \qquad 1{,}068 = \sqrt[2]{\frac{y_{2009}}{341989}}$$

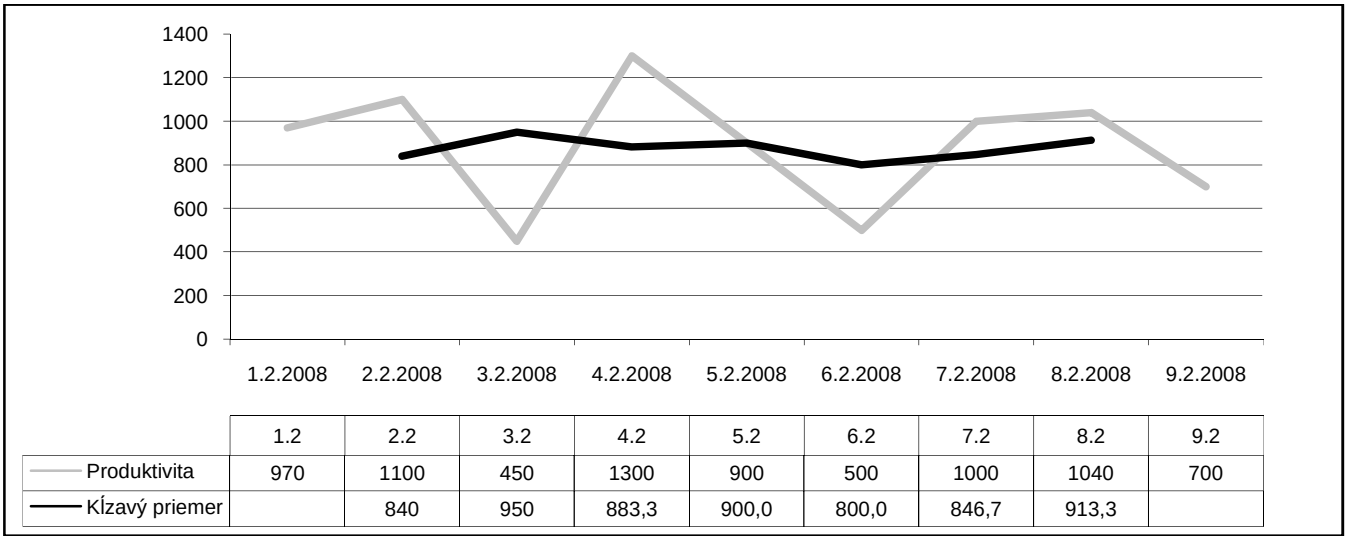
$$1{,}068 \cdot 341989 = y_{2008} \qquad 1{,}068^2 \cdot 341989 = y_{2009}$$

$$y_{2008} = 365\,244\ Kč / os. \qquad y_{2009} = 390\,081\ Kč / os.$$

8.1 Klouzavé průměry

Statistické sledování nějakého jevu nám obvykle přinesou haldu dat v časové řadě. Abychom si uměli z množství čísel vyvodit nějaký závěr, často je nutné **zvýraznit v časové řadě trend**, to znamená přestat vnímat odchylky mezi naměřenými hodnotami, ale soustředit se na to, jestli v globálu naměřená veličina roste nebo klesá a jak moc. Uvedu konkrétní příklad. Ve stavební firmě měří produktivitu práce tak, že každý den zaznamenávají množství vykonané práce do deníku a ohodnocují jej na základě prodejní ceny. Takže každý den na dané stavbě zaznamenají množství vykonané práce v penězích na osobu. Problémem je, že práce se často jeden den nejprve připraví – vykonají se činnosti, které mají minimální, pokud vůbec nějakou, prodejní cenu (stavba se zaměří, navozí se cihly, připraví se na místo, dovezou se stroje, písek a podobně) a na druhý den se vykoná samotná práce (vyzdí se stěna). Křivka produktivity nám tak skáče nahoru dolů a těžko říct, jaký je vlastně její trend, zda produktivita klesá, roste, nebo se drží na stejné úrovni.

Výborným nástrojem na vyrovnání takové křivky je klouzavý průměr. Jedná se o vyrovnání křivky tím způsobem, že **sčítáme první, druhou a třetí hodnotu, kterou jsme naměřili a výsledek vydělíme třemi**. To, co nám vyšlo je prostý klouzavý průměr pro druhou hodnotu $\bar{y}_2$. Obdobně, kdybychom chtěli vypočítat průměr pro třetí hodnotu $\bar{y}_3$, sčítáme druhou, třetí a čtvrtou hodnotu a vydělíme třemi. Klouzavý průměr se to jmenuje i proto, že jsme přeskočili první hodnotu, protože před ní není žádná, že které bychom vyrobili průměr. To, co je spočítané v grafu níže a co jsem tu popisoval, je **klouzavý průměr třetího řádu**. To proto, že průměrujeme tři čísla. Mohli bychom průměrovat i pět nebo sedm čísel – v případě sedmi bychom mohli začít s jeho počítáním pro čtvrtou hodnotu (1+2+3 **+ 4 +** 5+6+7), abychom si na obou stranách měli z čeho počítat průměr. A skončit bychom samozřejmě museli na $n-4$té hodnotě od konce.

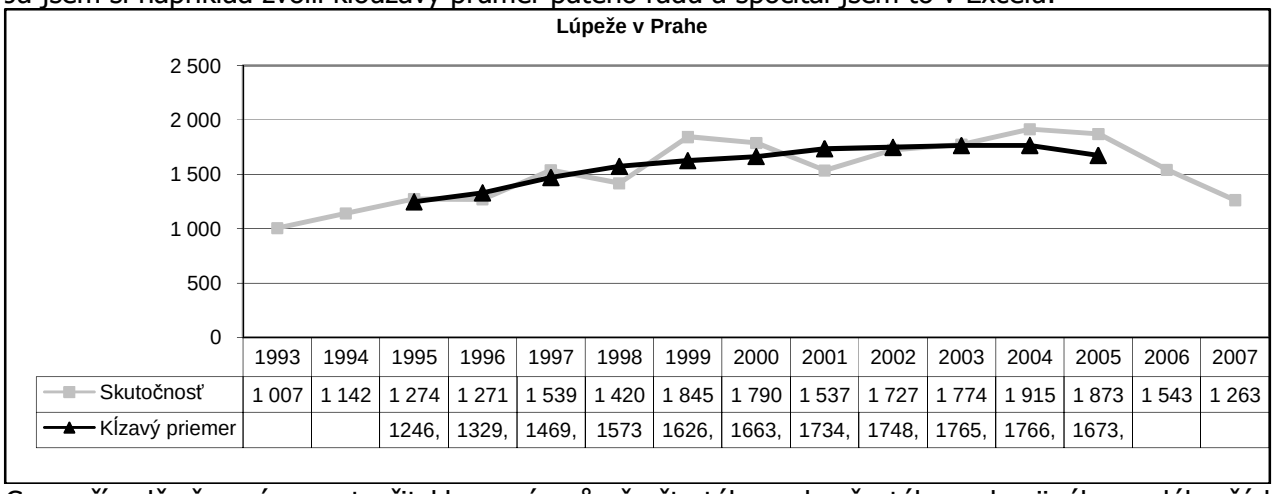


	1.2	2.2	3.2	4.2	5.2	6.2	7.2	8.2	9.2
Produktivita	970	1100	450	1300	900	500	1000	1040	700
KÍzavý priemer		840	950	883,3	900,0	800,0	846,7	913,3	

3. Očistěte časovou řadu počtu loupeží v Praze za období 1993 – 2007 od sezónnosti pomocí klouzavých průměrů.

1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
1007	1142	1274	1271	1539	1420	1845	1790	1537	1727	1774	1915	1873	1543	1263

Já jsem si například zvolil klouzavý průměr pátého řádu a spočítal jsem to v Excelu:



	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Skutočnosť	1 007	1 142	1 274	1 271	1 539	1 420	1 845	1 790	1 537	1 727	1 774	1 915	1 873	1 543	1 263
KÍzavý priemer			1246,	1329,	1469,	1573	1626,	1663,	1734,	1748,	1765,	1766,	1673,		

Co v případě, že máme vytvořit klouzavý průměr čtvrtého nebo šestého nebo jiného sudého řádu? V případě klouzavého průměru čtvrtého řádu bychom měli vzít první čtyři čísla časové řady, avšak ke které hodnotě napíšeme náš výsledek, ke druhé nebo třetí?

(45 + 50 + 34 + 39) / 4 = 42
Tady by teoreticky měl být výsledek

45	50	34	39

Teoreticky by to místo mělo být někde mezi druhou a třetí hodnotou – takže na pozici 2,5 – realita je taková, že to napíšeme ke třetí hodnotě. Obdobně u šestého řádu připíšeme náš klouzavý průměr ke čtvrté hodnotě a tak dále. Existuje však způsob, jak vyřešit i tento problém, a tím je **centrovaný klouzavý průměr**. Musíme se uvědomit, že klouzavý průměr čtvrtého řádu od hodnoty 1 po hodnotu 4 by nám nejlépe seděl někam na místo 2,5, a od hodnoty 2 po hodnotu 5 bychom jej umístili někam na 3,5. Proto stačí, když vezmeme tyto dva klouzavé průměry, sečteme je a vydělíme dvěma. A máme centrovaný klouzavý průměr čtvrtého řádu, který můžeme bez výčitek svědomí napsat ke třetí hodnotě časové řady a pokračovat dále v jeho výpočtu pro ostatní hodnoty.

8.2 Sezónní indexy

V případě, že máme časovou řadu vyrovnanou například klouzavými průměry, můžeme vypočítat i **empirické sezónní indexy**. Není to žádná věda, v podstatě jde jen o to, že **hodnotu v časové řadě vydělíme příslušnou hodnotou klouzavého průměru**. Pozor, v případě, že se chce, abychom počítali klouzavé průměry sudého řádu, musíme při počítání empirického sezónního indexu používat centrovaný klouzavý průměr. Pro třetí kvartál je to tedy 28 259 / 26 783 = 1,06.

4. *V tabulce jsou uvedené tržby v oblasti ubytování a stravování v ČR za období 2004 – 2007, vypočítejte empirické sezónní indexy a zdůvodněte, jak se projevuje sezónnost odvětví.*

Rok	**Q.**	**Tržby v Kč**	**Klouzavý průměr 4. řádu**	**Centrovaný klouzavý průměr 4. řádu**	**Empirický sezónní index**
2004	I	23 567			
	II	29 941			
	III	28 259	26 875	26 783	1,06
	IV	25 731	26 691	26 586	0,97
2005	I	22 832	26 481	26 469	0,86
	II	29 103	26 456	26 528	1,10
	III	28 160	26 601	26 705	1,05
	IV	26 307	26 809	26 985	0,97
2006	I	23 664	27 161	27 350	0,87
	II	30 512	27 538	27 831	1,10
	III	29 669	28 123	28 278	1,05
	IV	28 648	28 432	28 578	1,00
2007	I	24 897	28 725	28 845	0,86
	II	31 684	28 966	29 164	1,09
	III	30 633	29 363		
	IV	30 238			

Průměrný sezónní index	
I	0,86
II	1,09
III	1,05
IV	0,98

Sezónnost odvětví se projevuje v **průměrném sezónním indexu**, který dostaneme tak, že sečteme hodnoty empirických indexů za stejná čtvrtletí ve všech letech a vydělíme je počtem (zprůměrujeme je) – tak nám vyjde průměrný empirický index za každé čtvrtletí. Samozřejmě nezapočítáváme do tohoto průměru prázdná políčka, která vznikla z důvodu používání klouzavého průměru. Ve výsledku vidíme, že sezónní index má vyšší hodnoty ve druhém a třetím čtvrtletí – tedy v létě, kdy je ubytování a stravování atraktivnější než v zimním čtvrtletí. To potvrzuje sezónnost odvětví.

Jak jsem vzpomínal na začátku, žádaným výsledkem práce s časovou řadou je často určení trendu, neboli nějakého univerzálního pravidla, rovnice, podle které bychom mohli určit hodnotu časové řady v jakékoli chvíli v minulosti či budoucnosti. Trendy mohou být různé, představme si například cenu auta, ta má lineární trend, každým rokem, o který je auto starší, klesá i jeho cena, ale určitou hodnotu má stále. Rovnicí takového lineárního trendu je:

$$Y = \beta_0 + t \cdot \beta_1$$

Je to v podstatě regresní přímka a i její parametry jsou stejné, jako jsem je popsal v předchozí kapitole. Trendová křivka by však mohla být i parabola nebo exponenciála, ale i úplně jiná křivka, tak podle čeho vybrat tu správnou? Přímka je správná tehdy, když co nejlépe popisuje realitu, a to zjistíme například tak, že porovnáme v každém bodě časové řady hodnotu, která je naměřená, skutečná a hodnotu, která by nám vyšla použitím vzorce dané přímky. Pravděpodobně tam nějaký rozdíl bude, a to buď do plusu, nebo do mínusu. Protože nás nezajímají znaménka, a při sčítání minusových a plusových čísel by se nám navzájem vymazávali, raději ty odchylky umocníme na druhou, čímž odstraníme minusová čísla, ale velikost odchylek zůstane zachovaná (i když na druhou). Pokud spočítáme všechny tyto odchylky na druhou, vyjde nám hodnota, pomocí které se dá zhodnotit vlastnost použité přímky a nazývá se **reziduální součet čtverců**. Čím je jeho hodnota nižší, tím jsou i odchylky menší a křivka s menší hodnotou reziduálního součtu čtverců je tím pádem i vhodnější. Dalším koeficiente na určení vhodnosti použité křivky je **index determinace:**

$$R^2 = \frac{\sum (Y_t - \overline{y})^2}{\sum (y_t - \overline{y})^2},$$

který dává do poměru:

- součet druhých mocnin rozdílů mezi vypočítanou hodnotou a průměrem celé řady
- a součet druhých mocnin rozdílů mezi hodnotou řady v každém bodě a jeho průměrem

Tento koeficient udává kvalitu regresního modelu – neboli říká nám, zda je použitá křivka vhodná nebo ne. Ještě přesněji – udává, kolik procent rozptylu sledované proměnné (v našem případě hodnoty časové řady) je vysvětlených použitou křivkou. Tuto hodnotu vyjadřuje v procentech, pohybuje se tedy v intervalu <0;1>. Samozřejmě, čím větší hodnota, tím víc je zvolená křivka pro danou časovou řadu vhodná. Index determinace má však jeden problém, který snižuje jeho kvalitu, a to takový, že jeho velikost je ovlivněná počtem parametrů, které má křivka, jejíž vhodnost zkoumáme. Větší počet parametrů automaticky zvyšuje jeho hodnotu (zatímco lineární trend má dva parametry, parabolický má tři a tak dále), což v konečném důsledku může vést k preferenci složitějších křivek navzdory tomu, že reálně by byla vhodnější křivka jednodušší. Tato nepříjemnost se však dá odstranit, a to použitím **modifikovaného indexu determinace**:

$$R^2_{adj} = 1 - (1 - R^2) \cdot \frac{n-1}{n-p}$$

Přičemž n je počet bodů časové řady a p je počet parametrů křivky, jejíž vhodnost zkoumáme. Index determinace nebo lépe jeho modifikovanou hodnotu používáme obvykle při hodnocení vhodnosti křivky (např. na základě výstupu z programu SAS).

8.3 Regresní přístup k sezónní složce

Nechť máme data, která jsou sezónní, jako například údaje o velikosti HDP, které je většinou prvního čtvrt roku nižší než jiné kvartály. Použitím regresního přístupu dostaneme rovnici, která se bude skládat:

1. z **trendové složky**, která nám bude říkat, jak by se veličina vyvíjela nebýt sezónních výkyvů. V případě lineárního trendu bude její rovnice $\beta_0 + t \cdot \beta_1$.
2. ze **sezónních faktorů**, což je vlastně hodnota pro každý čtvrtrok, kterou odečteme anebo přičteme k hodnotě, kterou udává trendová složka.

Pravidlo je, že sezónní faktory, které odečteme anebo přičteme k jednotlivým čtvrtletím, musí v součtu za jeden rok (4 čtvrtletí) činit nulu, aby se nám nezvýšila ani nesnížila hodnota zkoumané veličiny za rok. Toto pravidlo se musí dodržet, má-li časový řad aditivní sezónnost. To znamená, že sezónnost se každý rok opakuje více či méně stejně (např. každý první čtvrtletí je HDP přibližně o stejnou částku nižší). V případě lineárního trendu bude rovnice, kterou opisujeme průběh křivky taková:

$$y_t = \beta_0 + \beta_1 \cdot t + \alpha_1 x_{1t} + \alpha_2 x_{2t} + \alpha_3 x_{3t} + \alpha_4 x_{4t}$$

$\beta_0 + t \cdot \beta_1$ je hodnota trendové složky, proměnné alfa jsou hodnoty složek pro první a druhý čtvrtletí a proměnné x nabývají hodnotu buď 1, nebo 0. To je dobré nato, že když máme první čtvrtletí, proměnná x_{1t} má hodnotu 1 a ostatní x jsou 0 – to znamená, že trendovou funkcí udanou hodnotu pro první kvartál, upravíme o sezónní hodnotu pro první čtvrtletí určenou parametrem alfa a zbylé alfy nás nezajímají. Podobně v druhém čtvrtletí má proměnná x_{2t}, hodnotu 1 a zbylé x jsou 0.

5. *Pro první až čtvrtý kvartál v letech 2004 až 2006 byl na základě kvartálních údajů o počtu zákazníků cestovní kanceláře vypočítán pomocí regresního přístupu následující výstup:*

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > ItI
Intercept	1	2,50000	0,31497	7,94	<0,001
t	1	0,31250	0,03050	10,25	<0,001
x1t	1	5,60417	0,29626	18,92	<0,001
x2t	1	-0,37500	0,28825	-1,30 (x 4,30)	<0,2343
x3t	1	6,97917	0,28336	24,63	< 0,001

   a. *napište rovnici trendové funkce*
   b. *vypočítejte sezónní faktory pro jednotlivé čtvrtroky a uveďte, v jakých měrných jednotkách jsou uvedené*
   c. *uveďte předpoklad pro třetí čtvrtletí 2007*

Uvedené hodnoty jsou výstup ze SASu, ve kterém jsou vypočítané odhadované parametry (parametr estimate) trendové a sezónní složky daného jevu. Intercept je odhad pro parametr β_0, téčko je odhad parametru β_1, a **pozor**, x1t, x2t a x3t jsou odhady parametru α jedna až tři. Kdybychom nebrali ohled na sezónnost, odhady by nám přímo udávaly hodnotu trendové funkce. Je proto nutné ještě upravit funkci $2,5+0,3125t$ podle sezónních parametrů. Sezónní parametry musíme zase přepočítat tak, abychom splnili podmínku jejich nulového součtu. Prvním krokem je výpočet hodnoty jejich průměru.

$$\overline{a}=\frac{a_1+a_2+a_3}{4}=\frac{5,60417-0,375+6,97917}{4}=3,052085$$

Ve jmenovateli musí být čtyřka, protože máme čtyři kvartály. V čitateli máme sice jen tři proměnné, ale právě takovým postupem získáme nulový konečný součet sezónních složek. Vypočítáme tedy průměr, upravíme každý sezónní faktor o vypočítaný průměr a ten čtvrtý, který jsme neměli v čitateli, se bude rovnat záporné hodnotě vypočítaného průměru. Dalším krokem je výpočet samotné sezónní hodnoty pro každý kvartál.

$$S_1=a_1-\overline{a}=5,60417-3,052085=2,552085$$
$$S_2=a_2-\overline{a}=-0,375-3,052085=-3,427085$$
$$S_3=a_3-\overline{a}=6,97917-3,052085=3,927085$$
$$S_4=-\overline{a}=-3,052085$$
$$2,552085-3,427085+3,927085-3,052085=0$$

Sezónní faktory nyní činní dohromady nulu. Teď můžeme vypočítat hodnotu nové trendové křivky, přičemž do ní zakomponujeme průměrnou kvartální výšku trendu podle rovnice:

$$T_t=(b_0+\overline{a})+b_1t=(2,5+3,052085)+0,3125\cdot t=5,552085+0,3125\cdot t$$

Výsledná rovnice, obohacená ještě o sezónní úbytky či přírůstky, se kterou budeme schopni předpovídat budoucí hodnoty časového řadu je:

$$Y_t=5,552085+0,3125t+2,552085x_{1t}-3,427085x_{2t}+3,927085x_{3t}-3,052085x_{4t}$$

Pro ty, co to nepochopili, to zkusím vysvětlit ještě jednou polopaticky. Co jsme vlastně spravili? Měli jsme křivku, která byla sezónní – v prvním a třetím kvartálu měla cestovní kancelář víc zákazníků než je obvyklé – neboť je dobrý čas pro dovolenou, a naopak v druhém a čtvrtém kvartálu jich měla méně. Ze zadání příkladu jsme hned mohli zaznamenat trendovou funkci. Ale pozor, ta funkce, kterou jsme zaznamenali, vypovídá jen o počtu té části zákazníků, kteří jsou stálými, skutečný počet zákazníků je vyšší. Například nechť jsou počty zákazníků za jednotlivé kvartály: 1Q-1500 2Q-800 3Q-1700 4Q-1000. Ze zadání bychom zaznamenali trendovou funkci, která by odhadovala 1000 zákazníků za kvartál, což je 4000 za rok. V skutečnosti jich máme 1500+800+1700+1000=5000. Z toho vyplývá, že chybějící tisícka je sezónní složka, kterou nemáme přičtenou. Proto musíme vyčíslit průměrnou sezónní složku, kterou zakomponujeme do trendové funkce (například nám chybí tisícka, takže bychom trendové složce přidali 1000/4=250 navíc ke každému kvartálu). Když máme takto upravený trend (1250 zákazníků za kvartál), který nám už ukazuje reálné počty zákazníků za rok, potřebujeme do něho dostat sezónnost. Je jasné, že trendová složka už obsahuje všechny zákazníky (za rok máme 5000, trendová funkce napočítá také 5000), proto nám sezónní faktory nemůžou snížit ani zvýšit

celkový počet zákazníků za rok. Už mohou jen v jednom kvartálu ubrat a v jiném přidat, ale přírůstky i úbytky musí dohromady dávat nulu. Když jsme takovéto sezónní složky vypočítali, můžeme sestavit finální rovnici, která nám umožní předpovědět vývoj v budoucnosti.
Hodnoty parametrů jsou v takových jednotkách, v jakých je veličina, kterou počítáme. Kdybychom počítali v tisíckách zákazníků, tak jsou to tisícky zákazníků. Kdybychom nyní chtěli spočítat jakoukoliv hodnotu, bylo by to možné. Nejdřív bychom ale museli vědět, které období má číslo jedna. Když by bylo řečeno, že obdobím jedna je například první kvartál 2004, výpočet hodnoty pro třetí kvartál 2007 by spočíval jen v zjištění pořadového čísla období a v jeho dosazení do vzorce. 3Q 2007 je patnáctým obdobím, proto za t dosadíme 15 a všechny proměnné x budeme považovat za nuly, jenom x_3 za jedničku (třetí kvartál).

$$Y_{15} = 5,552085 + 0,3125 \cdot 15 + 2,552085 \cdot 0 - 3,427085 \cdot 0 + 3,927085 \cdot 1 - 3,052085 \cdot 0 = \underline{14,16667}$$

9 Indexy

9.1 Jednoduché indexy

Slovo index má více významů a zatím co v databázích nebo knihách označuje nástroj pro rychlé vyhledávání konkrétního obsahu – registr, v matematice a mnoha dalších oblastech se jedná hlavně **o poměr něčeho**. Jednoduše řečeno, **jedno číslo vydělím druhým a mám z toho index**. Například si spočítám, kolikrát jsem byl tento týden (týden č. 1) na WC a vyjde mi nějaká hodnota. Stejně budu zaznamenávat hodnoty z dalších týdnů č. 2, 3,… Vydělíme-li tyto vypočítané hodnoty hodnotou z týdne č. 1, dostanu *„index chodění na WC"*. Dokonce tento můj vymyšlený index není jakýkoliv, ale je to **jednoduchý bazický index**, protože všechny hodnoty udávám jako poměr k první naměřené hodnotě. V případě zadání několika konkrétních čísel by to vypadalo takto:

Týden	1	2	3	4	5	6	7
Počet návštěv WC	20	16	24	22	38	23	21
Bazický index WCI (WC index)	1	0,8	1,2	1,1	1,9	1,15	1,05

Bazické indexy se tedy počítají jako:

$$\frac{q_2}{q_1};\frac{q_3}{q_1};\frac{q_4}{q_1};\ldots;\frac{q_s}{q_1}$$

Jinou formou jednoduchého indexu je **řetězový index**, který zase staví na tom, že hodnotu indexu pro každé období nestanoví na základě prvního období, ale na základě období předcházejícího. Kdybychom se drželi našeho indexu WCI, tak bychom museli dávat do poměru počet návštěv WC za x-tý týden s počtem návštěv za x-1 týden. V tomto konkrétním případě pro nás nemá řetězová forma indexu WCI žádnou vysokou vypovídací hodnotu. Jiné by bylo, kdybychom chtěli, aby se hodnota sledovaného parametru neustále zvyšovala. V tomto případě by bylo pro nás velice zajímavé zjistit, jaký je nárůst oproti předešlému období. I když nemáme zájem o kontinuální zvyšování hodnoty sledované veličiny v řetězovém indexu WCI, spočítat ho můžeme.

Týden	1	2	3	4	5	6	7
Počet návštěv WC	20	16	24	22	38	23	21
Řetězový index WCI (WC index)	-	0,8	1,5	0,92	1,73	0,6	0,91

Řetězový index je tedy vyjádřený jako:

$$\frac{q_2}{q_1};\frac{q_3}{q_2};\frac{q_4}{q_3};\ldots;\frac{q_s}{q_{s-1}}$$

Pro první období jeho hodnotu neuvádíme, začínáme až druhým obdobím.

1. Doplňte do tabulky chybějící údaje:

	Řetězový index	Bazický index
2000		1
2001	1,03	
2002	1,01	
2003		1,1
2004	1,1	
2005	0,9	
2006		1,1

Řetězový index nemá první hodnotu. Pro vztah řetězového a bazického indexu platí trojčlenka:

$$\text{řetězový index roku } x = \frac{\text{bazický index roku } x}{\text{bazický index předešlého roku}} \quad anebo: \; Ir_x = \frac{Ib_x}{Ib_{x-1}}$$

Když chceme získat bazický index 2001 a známe přitom bazický index 2000 a řetězový index 2001, tak stačí dosadit čísla a pak jen dopočítat výsledek. Pomocí této trojčlenky jednoduše vypočítáme řádek po řádku. Výsledek (ze shora dolů): - ; 1,03 ; 1,0403 ; 1,0574 ; 1,21 ; 1,089 ; 1,01

V celku vtipným příkladem indexu je tzv. Big Mac index. Ten porovnává cenu Big Macu v různých krajinách světa na základě teorie cenové parity, která říká, že cena by měla být po kurzovním přepočtu stejná. Index ukazuje, jak se liší cena v jednotlivých státech. Když například v ČR vychází cena Big Macu v dolarech vyšší než v USA, můžeme konstatovat, že česká měna je vůči americké nadhodnocena.

9.2 Souhrnné indexy

Doposud jsme hovořili o jednoduchých indexech, kdy jsme porovnávali jednu veličinu s tou stejnou akorát z jiného období. Situace se trochu změní, pokud budeme sledovat více veličin (například cenu aut a hrušek). Jednoduše nemůžeme sečíst takovéto věcně rozdílné veličiny. Zatímco u hrušky jde v ceně o halíře, u aut nehrají roli ani stokoruny. Jedná se tedy o **indexy, které se snaží vyjádřit změnu velikosti více sledovaných veličin**, které jsou ale natolik **odlišné povahou nebo měrnou jednotkou**, že je nemůže sčítat nebo zprůměrovat. Z tohoto důvodu používáme pro výpočet souhrnných indexů speciální vzorce. Protože se celý čas věnujeme ekonomické statistice, také tyto vzorce budeme používat na řešení ekonomických problémů. Konkrétně se jedná o tyto případy:

- **indexy úrovně** – sledujeme změnu ceny za určité období (kolik dnes stojí nákup skládající se z 10kg mouky, jednoho pracího prášku a zubního kartáčku a kolik stál stejný nákup před rokem)

- **indexy množství** – sledujeme změnu objemu výroby za určité období (o kolik více či méně se vyrobilo aut v únoru oproti lednu)

Oba druhy indexů se pokusím vysvětlit na příkladě.

2. *Máme zadané ceny a prodej maloobchodu, který prodává 4 druhy produktu ve dvou obdobích – leden (období 0) a únor (období 1). Pokusíme se pomocí souhrnných indexů vypočítat ze zadaných údajů všechno, co se dá.*

		Cena		Prodané množství	
	Výrobek	p0	p1	q0	q1
A	arašídy	50	55	200	190
B	banány	25	25	300	350
C	cukr	40	50	100	70
D	dvanáctka	13	10	220	250

Nejdříve se pokusíme vyjádřit souhrnnou změnu ceny, tj. jestli se produkt všeobecně zdražil, zlevnil, nebo zůstal stejný. První možností pro určení této veličiny je použití Laspeyresova indexu:

$$Ip^{(L)} = \frac{\sum_{i=1}^{n} p_{1,i} \cdot q_{0,i}}{\sum_{i=1}^{n} p_{0,i} \cdot q_{0,i}}$$

V čitateli násobíme únorovou cenou lednový objem prodeje daného výrobku a to z toho důvodu, abychom ocenili lednový prodej v únorových cenách. Nakonec jejich součiny sečteme. Ve jmenovateli násobíme to stejné pouze s tím rozdílem, že použijeme pouze lednové ceny. Tak nám v čitateli vyjde celá hodnota prodaného produktu v únorových cenách a ve jmenovateli ten stejný produkt ale v cenách lednových. Obě hodnoty tak můžeme porovnat.

$$Ip^{(L)} = \frac{\sum_{i=1}^{n} p_{1,i} \cdot q_{0,i}}{\sum_{i=1}^{n} p_{0,i} \cdot q_{0,i}} = \frac{55 \cdot 200 + 25 \cdot 300 + 50 \cdot 100 + 10 \cdot 220}{50 \cdot 200 + 25 \cdot 300 + 40 \cdot 100 + 13 \cdot 220} = \frac{25\,700}{24\,360} = 1,055$$

Výsledek Laspeyresova indexu nám říká, že produkt je v únoru o 5,5 % dražší než v lednu.

Další možností pro výpočet je použití Paascheho indexu. Na rozdíl od Laspeyresova indexu nesrovnává cenu minulého a současného období na základě objemu prodeje z minulého období, ale na základě objemu prodeje z běžného období, tj. namísto q0 používá q1.

$$Ip^{(P)} = \frac{\sum_{i=1}^{n} p_{1,i} \cdot q_{1,i}}{\sum_{i=1}^{n} p_{0,i} \cdot q_{1,i}}$$

Výpočet indexu potom probíhá podobně jako u Laspeyresa:

$$Ip^{(P)} = \frac{\sum_{i=1}^{n} p_{1,i} \cdot q_{1,i}}{\sum_{i=1}^{n} p_{0,i} \cdot q_{1,i}} = \frac{55 \cdot 190 + 25 \cdot 350 + 50 \cdot 70 + 10 \cdot 250}{50 \cdot 190 + 25 \cdot 350 + 40 \cdot 70 + 13 \cdot 250} = \frac{25\,200}{24\,300} = 1,037$$

Podle Paascheho indexu se ceny vzrostly o 3,7%. Těžko říci, který z těchto indexů je lepší, přesnější, nebo správnější. Z tohoto důvodu se používá ještě jeden, který není ničím jiným než geometrickým průměrem těchto dvou. Nazývá se Fischerův index:

$$Ip^{(F)} = \sqrt{Ip^{(L)} \cdot Ip^{(P)}}$$

Pro náš případ by měl Fischerův index hodnotu:

$$Ip^{(F)} = \sqrt{Ip^{(L)} \cdot Ip^{(P)}} = \sqrt{1,055 \cdot 1,037} = 1,046$$

Chceme-li vyjádřit změnu v objemu prodeje v nějaké univerzální jednotce, nepožijeme kusy, kartóny nebo trsy, ale **peníze**. Abychom uměli vypočítat tuto změnu objemu prodeje mezi dvěma obdobími v penězích, musíme ji vyjádřit **ve stejných cenách v obou obdobích**. Kdybychom nepoužili stejné ceny, tak vypočítaný kladný rozdíl by nemusel být způsobený počtem prodaných kusů, ale naopak zvýšením ceny a zachováním množství prodeje. Změnu objemu prodeje (výroby) vypočítáme pomocí indexů, které se nazývají úplně stejně jako ty indexy úrovňové, akorát že jsou to indexy objemu – první byl Laspeyresův index a druhý Paascheho index – tentokrát však oba objemové.

$$Iq^{(L)} = \frac{\sum_{i=1}^{n} p_{0,i} \cdot q_{1,i}}{\sum_{i=1}^{n} p_{0,i} \cdot q_{0,i}} \qquad Iq^{(P)} = \frac{\sum_{i=1}^{n} p_{1,i} \cdot q_{1,i}}{\sum_{i=1}^{n} p_{1,i} \cdot q_{0,i}}$$

Jak vidíme, příliš se neliší, jen Laspeyresův oceňuje objem výroby v cenách minulého období a Paascheho v cenách běžného období. Pojďme si vypočítat jejich hodnoty pro náš příklad:

$$Iq^{(L)} = \frac{\sum_{i=1}^{n} p_{0,i} \cdot q_{1,i}}{\sum_{i=1}^{n} p_{0,i} \cdot q_{0,i}} = \frac{50 \cdot 190 + 25 \cdot 350 + 40 \cdot 70 + 13 \cdot 250}{50 \cdot 200 + 25 \cdot 300 + 40 \cdot 100 + 13 \cdot 220} = \frac{24\,300}{24\,360} = 0,9975$$

$$Iq^{(P)} = \frac{\sum_{i=1}^{n} p_{1,i} \cdot q_{1,i}}{\sum_{i=1}^{n} p_{1,i} \cdot q_{0,i}} = \frac{55 \cdot 190 + 25 \cdot 350 + 50 \cdot 70 + 10 \cdot 250}{55 \cdot 200 + 25 \cdot 300 + 50 \cdot 100 + 10 \cdot 220} = \frac{25\,200}{25\,700} = 0,9805$$

Laspeyresův index říká, že prodej v únoru je na úrovni 99,75% lednového prodeje a Paascheho index říká, že je to 98,75%. Stejně jako u úrovňových indexů, tak u indexů objemů, můžeme spočítat ještě Fischerův index, který je geometrickým průměrem Laspeyresovho a Paascheho indexu.

$$Iq^{(F)} = \sqrt{Ip^{(L)} \cdot Ip^{(P)}} = \sqrt{0,9975 \cdot 0,9805} = 0,989$$

3. *V tabulce máme údaje o prodeji výrobků v maloobchodě. O kolik více musel vynaložit kupující v běžném období oproti období předcházejícímu? Vyjádřete absolutně i relativně!*

Výrobek	Prodej v běžném období	Index cen
A	250 000	1,2
B	320 000	1,1
C	140 000	1,4
D	175 000	1,75

Máme zadaný prodej v běžném období, to je podstatné $p_1 \cdot q_1$. Dále máme zadaný jednoduchý index cen běžného období pro každý výrobek, který nám říká, o kolik vzrostly ceny každého ze čtyř výrobků oproti minulému období. V případě, že vydělíme prodej běžného období každého výrobku tímto indexem, dostaneme objem prodeje běžného období pro každý výrobek oceněný v cenách období předešlého, což je v podstatě $p_0 \cdot q_1$. Máme-li zadané $p_1 \cdot q_1$ a $p_0 \cdot q_1$ a potřebujeme zjistit cenu, můžeme pro určení relativního cenového rozdílu použít Paascheho úrovňový index. Použijeme jeho verzi, která zahrnuje (zadaný v tabulce) jednoduchý index cen I_p. Jedná jen o matematickou úpravu, neboť po dělení prodeje v běžném období tímto indexem I_p dostaneme $p_0 \cdot q_1$.

$$Ip^{(P)} = \frac{\sum_{i=1}^{n} p_{1,i} \cdot q_{1,i}}{\sum_{i=1}^{n} \frac{p_{1,i} \cdot q_{1,i}}{I_p}} = \frac{\sum_{i=1}^{n} p_{1,i} \cdot q_{1,i}}{\sum_{i=1}^{n} p_{0,i} \cdot q_{1,i}} = \frac{250\,000 + 320\,000 + 140\,000 + 175\,000}{\frac{250\,000}{1,2} + \frac{320\,000}{1,1} + \frac{140\,000}{1,4} + \frac{175\,000}{1,75}} = \frac{885\,000}{699\,242} = 1,266$$

Relativní rozdíl je vyjádřený hodnotou indexu 1,266. To znamená, že kupující musel vynaložit o 26,6 % více korun. Absolutní rozdíl je rozdílem tržeb v běžných cenách a v cenách minulého období – tedy $p_1 \cdot q_1 - p_0 \cdot q_0 = 885\,000 - 699\,242 = 185\,758\,Kč$.

4. *2. V tabulce jsou uvedeny objemy výroby produktů za leden a index pro únor vypočítaný na lednové bázi. Vypočítejte Fischerův index změny celkové produkce v běžném období a vyjádřete ji v procentech.*

Výrobek	Výroba v minulém období	Cena v minulém období	Cena v běžném období	Index změny objemu výroby
A	112 ks	120 Kč	122 Kč	0,8125
B	300 kg	55 Kč	50 Kč	1,4
C	59 litrů	30 Kč	30 Kč	1,05
D	255 ks	75 Kč	85 Kč	1

Prvním krokem bude vypočítání únorové produkce. Nechť máme množství za leden a bazický index, který byl vypočítaný podle vzorce $\frac{q_2}{q_1} = I$. Chceme-li se dopracovat k množství q_2 (únor = běžné období), stačí nám upravit vzorec a množství dopočítat jako $q_2 = I \cdot q_1$.

Výroba	Výroba v minulém období	Výroba v běžném období
A	112 ks	91 ks
B	300 kg	420 kg
C	59 litrů	61,95 l
D	255 ks	255 ks

Poté dopočítáme Laspeyresovy a Paascheho objemové indexy podle vzorců a nakonec Fischerův index jako geometrický průměr Laspeyresova a Paascheho indexu.

$$Iq^{(L)} = \frac{\sum_{i=1}^{n} p_{0,i} \cdot q_{1,i}}{\sum_{i=1}^{n} p_{0,i} \cdot q_{0,i}} = \frac{120 \cdot 91 + 55 \cdot 420 + 30 \cdot 61,95 + 75 \cdot 255}{120 \cdot 112 + 55 \cdot 300 + 30 \cdot 59 + 75 \cdot 255} = \frac{55003,5}{50835} = 1,082$$

$$Iq^{(P)} = \frac{\sum_{i=1}^{n} p_{1,i} \cdot q_{1,i}}{\sum_{i=1}^{n} p_{1,i} \cdot q_{0,i}} = \frac{122 \cdot 91 + 50 \cdot 420 + 30 \cdot 61,95 + 85 \cdot 255}{122 \cdot 112 + 50 \cdot 300 + 30 \cdot 59 + 85 \cdot 255} = \frac{55635,5}{52109} = 1,0677$$

$$Iq^{(F)} = \sqrt{Ip^{(L)} \cdot Ip^{(P)}} = \sqrt{1,082 \cdot 1,0677} = \underline{1,0748}$$

Změna objemu výroby vyjádřená Fischerovým objemovým indexem je procentuálně + 7,48%.

10 Příklady pro procvičení

Z 20 hodnot byl vypočítán průměr 13 a rozptyl 7. Vypočítejte upravený rozptyl a průměr, pokud přidáme dvě hodnoty: 15 a 18.

$$\bar{x}_0 = 13 \qquad s_x^2 = 7 \qquad x_{21} = 15 \qquad x_{22} = 18$$

$$\bar{x} = \frac{\sum x_i}{n} \qquad s_x^2 = \frac{\sum x_i^2 - \bar{x}^2}{n}$$

$$13 = \frac{\sum x_i}{20} \qquad 7 = \frac{\sum x_i^2 - 13^2}{20}$$

$$\sum x_i = 13 \cdot 20 = 260 \qquad \sum x_i^2 = 309$$

$$\bar{x}_1 = \frac{260 + 15 + 18}{22} = 13,31 \qquad s_x^2 = \frac{309 + 15^2 + 18^2 - 13,31^2}{22} = 32,63$$

V letech 1991-1995 vzrostl průměr o 40% a rozptyl klesl o 30%, v letech 1996-2000 klesl průměr o 10% a rozptyl vzrostl o 20%. Jak se změní variační koeficient v letech 1991-2000?

$$91-95 : \bar{x} + 40\% \Rightarrow \bar{x} = 1 \cdot 1,4 = 1,4 \qquad s_x^2 - 30\% \Rightarrow s_x^2 = 1 \cdot 0,7 = 0,7$$

$$96-00 : \bar{x} - 10\% \Rightarrow \bar{x} = 1,4 \cdot 0,9 = 1,26 \qquad s_x^2 + 20\% \Rightarrow s_x^2 = 0,7 \cdot 1,2 = 0,84$$

$$V_k(91-95) = \frac{s_x}{\bar{x}} = \frac{\sqrt{0,7}}{1,4} = 0,597$$

$$V_k(96-00) = \frac{s_x}{\bar{x}} = \frac{\sqrt{0,84}}{1,26} = 1,336$$

$$\Delta V_k = 1,336 - 0,597 = 0,739$$

Politická strana předpovídala, že získá 5% voličů. Byl proveden průzkum a zjistilo se, že danou stranu volilo 50 lidí z 1180. Můžeme potom tvrdit na hladině alfa = 0,1, že strana získá méně než 5% hlasů?

$$p = 50/1180 = 0,04237 \qquad \alpha = 0,1 \qquad \pi_0 = 0,05$$

$$H_0 : \pi = \pi_0 \qquad W_a = \{U \le -u_{1-\alpha}\}$$

$$H_1 : \pi < \pi_0 \qquad u_{0,9} = 1,282$$

$$U = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{n}}} = \frac{0,04237 - 0,05}{\sqrt{\frac{0,05 \cdot 0,95}{1180}}} = \frac{-0,00763}{0,0063446} = 1,202$$

$$1,202 \le -1,282 \Rightarrow \textit{neplatí}, \textit{zamítáme } H_1$$

Životnost pneumatiky, která je náhodnou veličinou, má střední hodnotu 28 500 km a směrodatní odchylku 1900 km. Vypočítejte pravděpodobnost, že náhodně vybraná pneumatika bude mít životnost větší než 30 000 km a pravděpodobnost, že všechny 4 pneumatiky budou ještě v pořádku po 28 000 km.

$$P(X > 30000) = 1 - P(X \le 30000) = 1 - \Phi\left(\frac{X-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{30000-28500}{1900}\right) = 1 - 0,78524 = 21,47\%$$

$$P(X > 28000) = 1 - P(X \le 28000) = 1 - \Phi\left(\frac{28000-28500}{1900}\right) = 1 - \Phi(-0,26) = 1 - (1 - \Phi(0,26)) = 60,25\%$$

$$P_1 \cap P_2 \cap P_3 \cap P_4 = 0,6025 \cdot 0,6025 \cdot 0,6025 \cdot 0,6025 = 13,17\%$$

Náhodně jsme vybrali 50 osob určitého zaměstnání a zjišťovali jsme jejich platy. Výběrový průměr vyšel 17286 a výběrová směrodatná odchylka 2722. Na hladině významnosti 0,05 ověřte hypotézu, že průměrný plat v základním souboru je 17520.

$$n = 50 \qquad \overline{x} = 17286 \qquad s'_x = 2722 \qquad \alpha = 0,05$$

$$H_0 : \mu_0 = 17520 \qquad U = \frac{\overline{x} - \mu_0}{s'_x} \cdot \sqrt{n}$$

$$H_1 : \mu_0 \neq 17520 \qquad W_a = \left\{ |U| \ge u_{1-\alpha/2} \right\}$$

$$U = \frac{17286 - 17520}{2722} \cdot \sqrt{50} = 0,6078$$

$$u_{0,975} = 1,96 \qquad 0,6078 \ge 1,96 \Rightarrow \textit{neplatí, zamítáme } H_1$$

Jeden pracovník splní daný úkol za 8,5 hodiny, druhý za 7 hodin a třetí za 6,8 hodiny. Jaká je průměrná doba plnění úkolu?

$$\overline{x}_H = \frac{n}{\frac{1}{\sum x_i}} = \frac{3}{\frac{1}{8,5} + \frac{1}{7} + \frac{1}{6,8}} = 7,36$$

Je dána tabulka ze SASu – parametry časové řady výroby piva v hektolitrech v čtvrtletích od roku 2005. Kolik se bude vyrábět piva ve 4 čtvrtletí 2006?

Variable	DF	Parameter Estimate
Intercept	1	5
t	1	0,12
x1t	1	-0,123
x2t	1	1,234
x3t	1	1,004

$$\overline{a} = \frac{a_1 + a_2 + a_3}{4} = \frac{-0,123 + 1,234 + 1,004}{4} = 0,52875$$

$$S_1 = a_1 - \overline{a} = -0,123 - 0,52875 = -0,65175$$

$$S_2 = a_2 - \overline{a} = 1,234 - 0,52875 = 0,70525$$

$$S_3 = a_3 - \overline{a} = 1,004 - 0,52875 = 0,47525$$

$$S_4 = -\overline{a} = -0,52875$$

$$-0,65175 + 0,70525 + 0,47525 - 0,52875 = 0$$

$$T_t = (b_0 + \bar{a}) + b_1 t = (5 + 0,52875) + 0,12 \cdot t = 5,52875 + 0,12 \cdot t$$

$$Y_t = 5,52875 + 0,12\,t - 0,65175\,x_{1t} + 0,70525\,x_{2t} + 0,47525\,x_{3t} - 0,52875\,x_{4t}$$
$$Y_8 = 5,52875 + 0,12 \cdot 8 - 0,52875 = 5,96$$

Ve velkém textilním podniku pracuje 70% žen. Provedeme náhodný výběr 50 pracovníku (s vracením). Zjistěte střední hodnotu a rozptyl.

$$s\,vracením \Rightarrow binomické\,rozdělení$$
$$E(X) = n \cdot \pi \qquad D(X) = n \cdot \pi(1-\pi)$$
$$E(X) = 50 \cdot 0,7 = 35$$
$$D(X) = 50 \cdot 0,7 \cdot (1-0,7) = 10,5$$

Pravděpodobnost, že se kufr ztratí v 1. letadle je 1%, pravděpodobnost, že se ztratí v druhém je 3% a ve třetím 2%. Jaká je pravděpodobnost, že se kufr stratí právě v i-tém letadle (i=1,2,3)?

$$P(1) = 0,01$$
$$P(2) = 0,99 \cdot 0,03 = 0,0297$$
$$P(3) = 0,99 \cdot 0,97 \cdot 0,02 = 0,019206$$

Vyberte vhodnější model a odůvodněte. Odhadněte y pro x=20. Počet n byl 8.

a. *přímka: -16,830-2,244x*
b. *parabola: 28,982-5,756x+0,305x²*

	St	Sr	Sy
Přímka	875,888	253,612	1129,5
Parabola	1025,7	103,804	1129,5

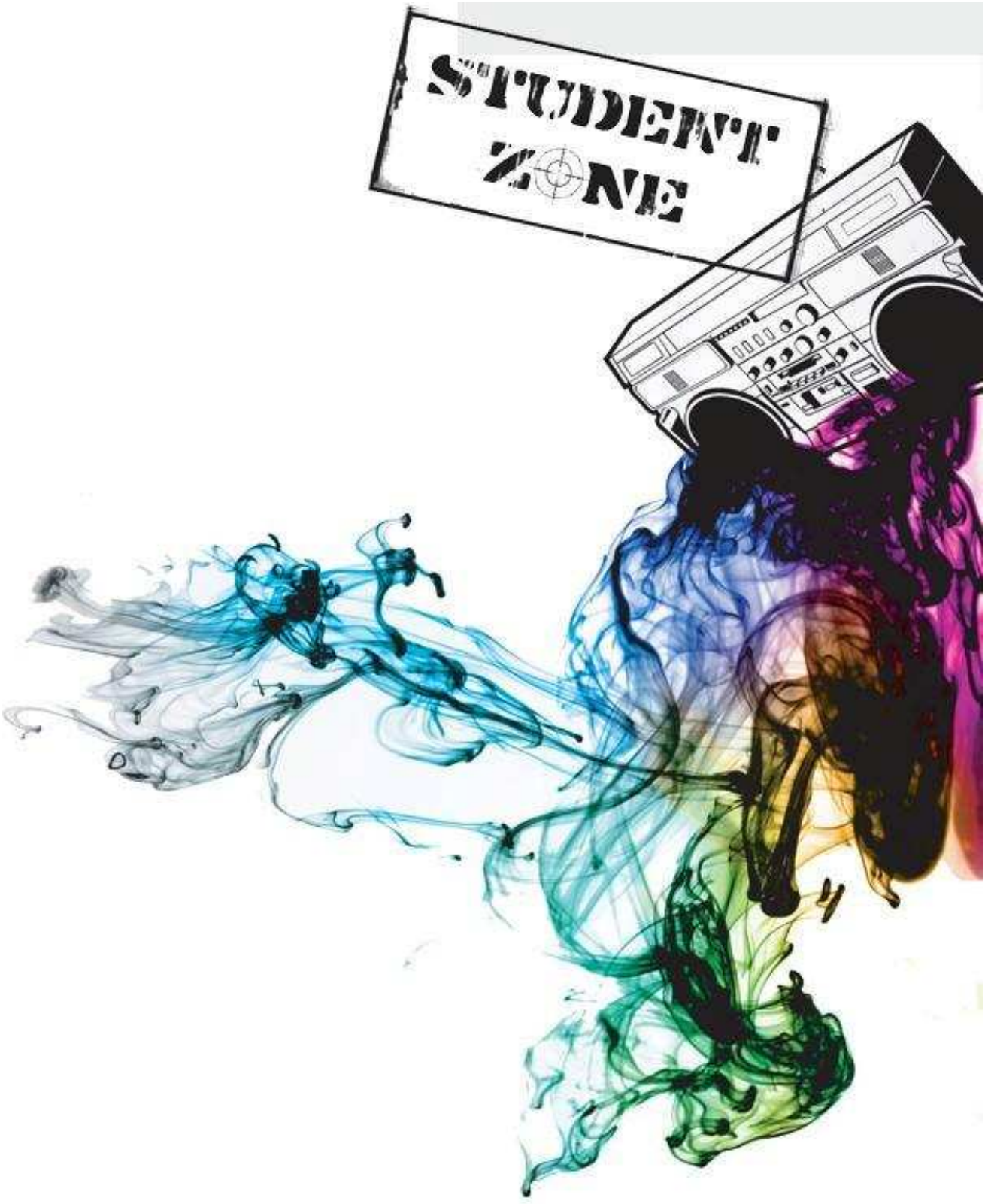
$$n = 8 \qquad R^2_{adj} = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-p} \qquad R^2 = \frac{S_T}{S_y}$$

$$R^2_{adj}(přímka) = 1 - \left(1 - \frac{875,888}{1129,5}\right) \cdot \frac{8-1}{8-2} = 1 - 0,225 \cdot 1,16 = 0,739$$

$$R^2_{adj}(parabola) = 1 - \left(1 - \frac{1025,7}{1129,5}\right) \cdot \frac{8-1}{8-3} = 1 - 0,092 \cdot 1,4 = 0,8712$$

Vhodnější je parabola.

$$y_{20} = 28,982 - 5,756 \cdot 20 + 0,305 \cdot 20^2 = 35,862$$


STUDENT
ZONE